UNIVERSITY OF CALIFORNIA
RIVERSIDE

Information Losses in Neural Classifiers With Applications to Training Data Selection Strategies and Cyber Physical Systems

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

by

Brandon James Foggo

December 2019

Dissertation Committee:
- Dr. Nanpeng Yu, Chairperson
- Dr. Yingboa Hua
- Dr. Christian Shelton

The Dissertation by Brandon James Foggo is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ABSTRACT OF THE DISSERTATION

Information Losses in Neural Classifiers With
Applications to Training Data Selection Strategies
and Cyber Physical Systems

by

Brandon James Foggo

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, December 2019
Dr. Nanpeng Yu, Chairperson

This dissertation considers the subject of information losses arising from finite datasets used in the training of neural classifiers. It proves a relationship between such losses and the product of the expected total variation of the estimated neural model with the information about the feature space contained in the hidden representation of that model. It then bounds this expected total variation as a function of the size of randomly sampled datasets in a fairly general setting, and without bringing in any additional dependence on model complexity. It ultimately obtains bounds on information losses that are less sensitive to input compression and much tighter than existing bounds. It then uses these bounds to explain some recent experimental findings of information compression in neural networks which cannot be

explained by previous work. The dissertation goes on to provide analytical derivations for the relationship between neural architectures and the mutual information contained in their representations, which can be useful for guided architecture selection schemes. It then uses these developments to propose and illustrate a new framework for analyzing training data selection methods. The dissertation use this framework to prove that facility location methods reduce these losses, and then derive a new data dependent bound on them. This bound can be used to evaluate datasets and acts as an additional analytical tool for the study of data selection techniques. The dissertation then applies this theory to the problem of Phase Identification in power distribution systems. In particular, it focuses on improving supervised learning accuracies by exploiting some of the problem's information theoretic properties. This focus, along with the advances developed earlier in this work, helps us create two new Phase Identification techniques. The first transforms the bound on information losses into a data selection technique. This is important because phase identification data labels are difficult to obtain in practice. The second interprets the properties of distribution systems in the terms of the information losses developed earlier in the dissertation. This allows us to obtain an improvement in the representation learned by any classifier applied to the problem. Furthermore, since many problems in cyber-physical systems share similarities to the physical properties of phase identification exploited in this dissertation, the techniques can be applied to a wide range of similar problems.

TABLE OF CONTENTS

List of Tables xi

List of Figures xii

Chapter 1: Introduction and Background 1

1.1 Introduction 1

Chapter 2: Background 8

2.1 Some Preliminaries 8

2.1.1 Information Theory 8

2.1.2 Estimation and Control of Mutual Information 9

2.1.3 Large Deviations Theory 12

2.2 Notation and Assumptions 13

2.3 Background 15

2.3.1 The Information Bottleneck Principle 15

2.3.2 Information and Generalization 15

2.3.3 Information Losses 16

2.3.4 Automatic Implementation via Neural Networks 18

2.3.5 Training Data Selection Related Works 19

2.3.6 Phase Identification Related Works 20
2.3.7 Battery Storage Literature Review 23
2.3.8 Battery Modelling 25
2.3.9 Degradation Modeling 28

Chapter 3: Bounds on Information Losses 31

3.1 Product Form Decomposition - Setup 31
3.1.1 Product Form Decomposition - Theorem and Proof 35
3.1.2 Understanding $\bar{\delta}(\hat{\mathbb{P}})$ 38
3.2 Finite Bounds for Discrete Spaces 39
3.2.1 Bounding $\bar{\delta}(\hat{\mathbb{P}})$ - Setting 42
3.2.2 Bounding $\bar{\delta}(\hat{\mathbb{P}})$ - The Asymptotic Case 45
3.2.3 Bounding $\bar{\delta}(\hat{\mathbb{P}})$ - The Non-Asymptotic Case 48
3.2.4 Some Insights 53
3.3 Experiments 54
3.3.1 How These Bounds Solve Experimental Discrepancy 54
3.3.2 Tightness of Bounds 54
3.4 Chapter Conclusion 58

Chapter 4: The Maximum Mutual Information of Varying Architectures 59

4.1 MMI Calculations 59
4.2 Single Layer Linear Networks (Fully Connected and Convolutional) 59

4.2.1 Fully Connected Case 59

4.2.2 Convolutional Case 66

4.3 Single Layer Relu Networks 68

4.3.1 Relu Activations 68

4.4 Multilayer Networks 75

Chapter 5: Information Losses in Training Data Selection Strategies 76

5.1 Facility Location Optimization Mitigates Information Losses - First Approach 76

5.1.1 Metric Facility Location 76

5.2 Facility Location Optimization Mitigates Information Losses - Second Approach 79

5.3 A Data Dependant Bound for Information Losses 81

5.3.1 Converting to multiple classes 90

5.4 Experiments 91

5.4.1 Correspondence Between Bound and Classification Accuracy 91

5.5 Chapter Summary 92

Chapter 6: Mitigating Information Losses in Practice 94

6.1 Phase identification Properties and Similar Problems 94

6.2 Technique - Inverse Schur Training Data Selection 95

6.2.1 Motivation - Field Testing 95

6.2.2 Training Data Selection via Information Losses 95

6.3 Technique - Information Loading 97

6.3.1 Voltage Data and Phase Identification 97
6.3.2 Properties of Voltage Data 98
6.3.3 Entropy Analysis 101
6.3.4 Maximum Mutual Information (MMI) Estimation: Further Evidence of the Low Entropy Feature Space Hypothesis 104
6.3.5 Information Loading 105
6.4 Experiments 106
6.4.1 Data 106
6.4.2 Preprocessing 108
6.4.3 Results 108
6.5 Conclusion 112

Chapter 7: Battery Storage Policy with Degradation Mitigation 114

7.1 Decoupled Degradation Valuation and Properties 114
7.2 Degradation Linearization 121
7.3 Results 131
7.3.1 Linearization Performance 131
7.3.2 Value Loss from Degradation 133
7.3.3 Value Recovery 134
7.3.4 Heuristics and Parameter Estimation 136
7.3.5 Optimal R 136
7.4 Chapter Summary 137

Chapter 8: Conclusion 138

References 149

LIST OF TABLES

2.1 Degradation function parameters for a Li-Ion Battery. 30

5.1 Metric Assumptions 77

6.1 Distribution Circuits Characteristics 107

6.2 Baseline Establishment 109

6.3 Proposed Techniques 109

6.4 Accuracy comparisons between the literature and the proposed method. . . 109

7.1 Battery size and economic parameters. 132

LIST OF FIGURES

1.1 Illustration of the Phase Identification Problem. Here, each T represents a transformer, and each combination of the letters A, B, and C refers to a phase connection type. 5

2.1 The classification model assumed in this dissertation. 14

2.2 The DP Thevenin Equivalent Model of a battery 26

2.3 Open Circuit Votlage with respect to State of Charge. Figure from [110]. . . 27

3.1 Bayesian Network describing the relationships between random variables in the proof of Theorem 1. 36

3.2 (left) New bounds on a low entropy feature space (right) Old bounds on the same space. (Bottom) New bounds on a high entropy feature space. 55

3.3 $(\bar{\delta}(\mathbb{P}_f) - \zeta)$ for several datasets. (Blue) True confidence interval, (Red) bound [Theorem 5]. 56

3.4 $I^{(1)}_{Loss}$ for MNIST over varying architectures. (Blue) True confidence interval, (Red) Information bound [Theorem 1]. 57

5.1 Caption 90

5.2 Classification errors against the data quality measure with a fixed training data size for varying datasets. Dataset and corresponding rbf γ value are indicated at the top of each plot. 92

5.3 MNIST data quality measure and classification error against training data size for several methods of training data selection. 93

6.1 String diagram representation of the information loading forward pass. $r : \mathcal{X} \to \mathcal{Z}$ is the representation function. $t : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ is the argument of the information estimator. 107

6.2 Learned representations on circuit II. (Left) random selection with no information loading. (Right) Targeted selection with information loading. 110

6.3 Learned representations on circuit V. (Left) random selection with no information loading. (Right) Targeted selection with information loading. 110

6.4 Learned representations on circuits I (left) and IV (right) with both techniques implemented. 111

7.1 Time series data of the SoC over a sample 24 hour window. 119

7.2 Scatterplot of mean normalized SoC vs. normalized DoD. 122

7.3 E_{max} at the beginning of each year for three optimization schemes. 131

7.4 Percent error in E_{max} after each year. 132

7.5 Net Present Value (NPV) of each year. 133

7.6 Regulation Up signal variance vs. price. The trend line is $y = 97.86x - 0.81$.134

CHAPTER 1
INTRODUCTION AND BACKGROUND

1.1 Introduction

An estimator is limited to the information that it has about the variable it's estimating. But this information is limited to what the estimator has seen from the samples training it. The full information of a random variable cannot be transferred to an estimator by finite samples - some information is lost. This dissertation analyzes such losses for neural network classifiers. Analyzing these losses can lead to improved architecture designs and training data selection strategies, and provide explanations for empirical results in machine learning theory.

The study of these loses as a tool for deep learning theory arose from the attempts to understand neural network behavior through the concept of an information bottleneck [96, 97]. This theory was later investigated both analytically [3] and experimentally [83, 92]. They are used, primarily, as an explanatory tool which can act as a supplement to classical statistical learning theory (CSLT), which typically fails to explain the success of deep learning models (for example, deep networks tend to perform better when they have *higher* VC dimension, while CSLT would predict the opposite). We will further discuss the utility of these losses in this dissertation, and we will denote this newly arising field of deep learning theory as information theoretic deep learning theory (ITDLT).

But this theory is still somewhat incomplete. The reader will find that reference [83] above actually contradicts the others - giving experimental evidence *against* some of the claims established in the earlier works. In particular, ITDLT, as it previously stood, would claim that neural networks should always act as a lossy compressor of the input data - a claim which arises from bounds on information losses that are exponential in the information

content of the final hidden layer of the network (while still being smaller than CSLT bounds for larger networks). But experiments show that this is only *sometimes* true. While compression does seem to always occur when using saturating activation functions, like sigmoid and tanh, compression in networks using linear and relu activation functions seems to be more nuanced.

But instead of abandoning ITDLT, we believe that the theory can be improved in such a way that it explains all of these experiments. Since most contrary evidence to the theory can be traced to those exponential bounds, we hypothesize that these bounds, while tighter than those of CSLT, are still not quite tight enough to account for every experiment. In this dissertation, we aim to derive bounds which are much tighter than the existing ones.

With these new bounds, we will be able to explain the experimental discrepancy found in the above literature, giving detail into why *some* situations yield neural network compression, even with relu activation functions, and others do not. For example, in the case of low entropy feature spaces, our bounds show that there is simply not enough information to lose such that compression is beneficial.

This will lead to a better understanding of the information relationships found in neural networks, and to a better understanding of neural networks in general. This better understanding will allow guided development of network architectures and other algorithms which are theoretically sound.

In one critical step to achieving these bounds, we decompose information losses as a product of a term that mostly depends on network architecture and a term that mostly depends on the training dataset used to train that architecture. This decomposition can thus be applied to network architecture design and training data selection strategies independently. These aspects of applying this theory will be the subject of future work.

Finally, while these new bounds are much tighter than both CSLT bounds and the old ITDLT bounds, and while they are capable of explaining all experiments in literature, we

will see experimentally that these bounds are fairly tighter than they needed to be to achieve our goals.

We can immediately apply this theory to two fields. The first is the field of Active Learning. Some sampled points are more useful to a classification problem than others. Thus we may not need many labels for adequate performance. But finding useful data points can be challenging. This dissertation provides a new information theoretic framework for analyzing data selection methods. The framework studies a quantity called information losses, which compares the information content of a random variable to the information content of a random variable approximated through finite samples of the original. Roughly speaking, this quantity tells us the minimal amount of information lost from the class variable upon sampling, and is connected to classification accuracy through several analytical links such as Fano's inequality.

We first provide an example of this framework's usage. In particular, we will provide a new proof of viability for facility location based data selection methods. The analysis that we perform brings to light some particular advantages of our framework. In particular, the method of proofs that come about under our framework tend to be very intuitive. This is because our framework links the information theoretic quantities under study to simpler quantities which are more closely related to standard mathematical analysis. Thus in writing a somewhat intuitive analysis proof, we get a theoretically justified information theoretic proof for free. These proofs, and by extension the full analysis, should bring useful insights into the methods being studied.

We then derive a new bound on information losses. This bound has several benefits. First, it is dataset dependent. It can thus act as a data quality measure to evaluate active learning methods. Second, it acts as a new analytical tool that we can use in evaluating training data selection techniques. The bound is extremely tight to experiment - in fact, it is an order of magnitude tighter than the current tightest known bound in literature. The

bound also has predictive power towards classification accuracy. This follows directly from the link between information losses and classification error, but we provide experimental evidence of this connection as well - obtaining a tight correspondence between this bound and classification error.

Finally, we can apply this framework to the field of Cyber-Physical-Systems. In particular, we can apply it to a known problem in Power Systems called the *Phase Idenitification Problem*. A power distribution circuit encompasses several components. It contains busses, powerlines, substations, regulators, transformers, and more. The physical arrangement of these components constitute the circuit's topology, which dictates much of the system's operation and planning. Over time, a circuit's topology will change. For example, a power outage may initiate a structural change such that the number of the effected customers is minimized. But the topology documentation will often not follow this change - they are only typically updated after major distribution expansion projects. As a result, there are long periods of time in which a circuit's topology is wrong, and this poses a serious problem. Power flow analysis, state estimation, and Volt-VAR control all depend on accurate topological information. When this is not available, the usefulness of those methods diminishes, and the system runs less effectively as a whole. Methods for faster documentation updates are necessary.

Critical to the application of topology identification is the subproblem of phase identification. This subproblem describes the composition of each powerline in the network. Typically, a primary distribution powerline is made up of the four fundamental lines A, B, C and n. The primary feeder, fed directly from the power substation, often consists of all four. However, at some point, a subset of these lines will be branched from the feeder. This change usually happens once along the path from the substation to any customer. As such, we define a customer's 'phase type' as the lines branched along that path. This is illustrated in Figure 1.1. Knowledge of these phase types is necessary for estimation of electrical

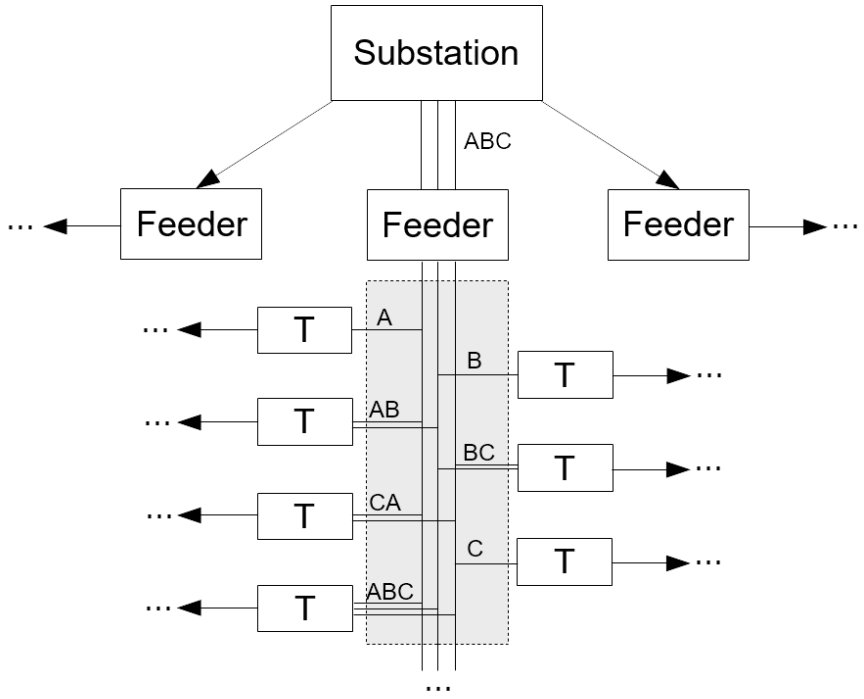

Figure 1.1: Illustration of the Phase Identification Problem. Here, each T represents a transformer, and each combination of the letters A, B, and C refers to a phase connection type.

distances, and for topology reconstruction in general. In fact, since topology estimation is often taken as a precursor to state and parameter estimation in distribution networks, we can think of phase identification as the first step in a pipeline of distribution system modelling techniques.

Literature on phase identification exists, but is limited. Like most power system applications, there are two broad classes of phase identification approaches - one model based, and the other data-driven. Like those other applications, model based methods have high interpretability but low accuracy, and data-driven methods have high accuracy but low interpretability. Our research poses itself as an intermediate between these extremes. This is done through Information Theoretic Machine Learning (ITML). This branch of machine learning connects traditional learning theory to the field of Information Theory. As such, interpretable notions such as entropy and mutual information are studied. These interpretable

measures can be connected to the physics of the problems that we wish to solve and used to our advantage.

We will focus on a particular regime of ITML which needs further development - the small data regime. Phase identification methods, particularly supervised methods, require a lot of training data to achieve high accuracies. But obtaining this training data requires portions of the lengthy field tests which we were trying to avoid in the first place. Thus we have a trade-off between phase identification accuracy and the amount of time that a distribution circuit's topology documentation is incorrect. In this dissertation, we will develop methods to obtain higher phase identification accuracies even when small training datasets are used.

In particular, we will develop two techniques: Inverse Schur Training Data Selection, and Information Loading. The first corresponds to selecting the most informative data-points prior to field testing. The second consists of a modification to the objective function of a standard learning algorithm, with some modifications to the training phase in order to support it. We will heavily emphasize the theoretical reasoning behind these techniques, particularly in why they are applicable to the phase identification problem. Our proposed techniques have the following benefits:

- They require little infrastructure or physical labor.
- They do not require any modeling of the network.
- They are robust with respect to missing data.
- They are easy to implement and tune.
- They can handle any variety of phase connections.
- They provide the best supervised accuracies to date.

- The representations learned are highly meaningful.
- They are generalizable to other networked systems.

Finally, we will discuss some contributions to Battery Storage Policy. Renewable generation sources are quickly penetrating the electric power grid. These sources are highly stochastic. As a result, renewable generation complicates the equalization of a network's power generation with its load. But when these values are not equal, the electrical frequency of that network deviates from its nominal value and this leads to complications throughout the network. The equalization can be made more reliable with Battery Energy Storage Systems (BESS). BESS have fast ramping rates and can dynamically switch between power generation and absorption to quickly offset any imbalance in a network's power generation and its load. However, BESS are expensive, and so the profitability of using BESS to mitigate network power imbalances remains under question.

BESS owners can obtain revenue streams via energy arbitrage and providing one or more of the ancillary services described in [50]. The main service investigated in this research is frequency regulation. The primary goal of this research is to maximize the revenue from these sources when realistic physical constraints are considered and all future prices are known. The latter assumption makes it so that the valuation obtained in this work is as an upper bound on actual BESS profitability.

Of critical importance is the lifetime of a BESS and how use of the system affects its longevity. A battery that undergoes frequent and powerful charging/discharging cycles will die sooner than a battery that doesn't. Thus long-term profitability depends on usage in a more complicated way than just revenue streams. This research establishes a method that optimizes profitability when lifetime and degradation are considered.

CHAPTER 2

BACKGROUND

2.1 Some Preliminaries

2.1.1 Information Theory

Information theory concerns properties of probability distributions similar to those seen in the fields of thermodynamics and statistical mechanics. Its primary quantity of concern is the Shannon entropy. Letting X be a random variable, this is computed via

$$H(p) = \mathbb{E}_X \left[log_2 \, p_X(X) \right] \tag{2.1}$$

Of central importance to this work are quantities relating pairs of random variables, and quantities comparing distributions over the same random variable. These are given by the mutual information:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = \mathbb{E}\left[log \, \frac{p_{XY}(X,Y)}{P_X(X)P_Y(Y)} \right] \tag{2.2}$$

and the Kullback-liebler divergence

$$\mathcal{D}_{KL}\left(P||Q\right) = \mathbb{E}_{X\sim P}\left[log \, \frac{P(X)}{Q(X)} \right] \tag{2.3}$$

respectively.

These quantities are useful for the following reasons. First, the mutual information between two variables tells us how much 'information' must be gathered to fully know the value of Y given that we know the value of X (on average). Thus, it tells us how much we

can learn about one variable from another. This is important to machine learning, as the central task of this field is to infer random variables (e.g. a class variable) from other random variables (e.g. a feature variable or a representation variable). Second, the kullback-liebler divergence is useful in bounding the probability of inferring a 'bad' distribution Q from samples of data generated by P. This will be more clear when we discuss the theory of large deviations in a latter section of this chapter.

2.1.2 Estimation and Control of Mutual Information

The bounds derived in this dissertation are less sensitive to $I(X;Z)$ than previous bounds. But the dependence is still there. In many cases, accuracy can still be gained by limiting the information present between X and Z. Even in cases where generalization accuracy cannot be gained by these limits, it may still be desirable to estimate $I(X;Z)$ and $I(Y;Z)$. For example, one may wish to visualize the evolution of these mutual informations as a neural network trains as was done by the authors of [92]. Here we will give a review of information estimators/controllers. We have divided the methods into two groups - methods which act like variational Inference (VI), and methods which act like Generative Adversarial Networks (GAN).

VI - like methods

Several methods of limiting $I(X;Z)$ have been proposed. Authors of [3] found a tight relationship between $I(X;Z)$ and $I(W;\mathcal{D}^l)$ where W is a random vector of neural network weights. They then noted that

$$\begin{aligned}\mathbb{E}\left[log_2\frac{p(w|\mathcal{D}^l)}{q(w)}\right] &= \mathbb{E}[log_2\frac{p(w|\mathcal{D}^l)}{p(w)}] + \mathbb{E}[log_2\frac{p(w)}{q(w)}] \\ &= I(W;\mathcal{D}^l) + \mathbb{E}[log_2\frac{p(w)}{q(w)}] \end{aligned} \tag{2.4}$$

where $q(w)$ is any *assumed* marginal distribution on the weights. Thus regularizing the KL divergence term $\mathbb{E}\left[log_2\frac{p(w|\mathcal{D}^l)}{q(w)}\right]$ will lead to regularization of $I(W;\mathcal{D}^l)$. This term is the same as that used in variational inference (e.g. Bayesian Neural Networks [57]). Thus methods such as variational dropout [53] [67] may be used. Unfortunately, this has two drawbacks. First, it suffers from the standard problem present in variational inference. That is, we must choose weight distributions $p(w|\mathcal{D})$ and $q(w)$ that lead to a tractable KL divergence. This limits our search space for the random variable Z. Second, we do not get this information regularization for free. Instead, it comes coupled with a second regularization term $\mathbb{E}[log_2\frac{p(w)}{q(w)}]$. This term penalizes any distribution whose marginal distribution differs from the assumed one. Both of these drawbacks will lead to further sub-optimality of our optimized variable Z.

A similar method called Information Dropout [2] regularizes

$$\mathbb{E}\left[log_2\frac{p(z|y)}{q(z)}\right] = I(X;Z) + \mathbb{E}[log_2\frac{p(z)}{q(z)}] \tag{2.5}$$

where $q(z)$ is again an assumed marginal distribution. Regularization of this term leads to methods similar to Variational Autoencoders [54]. These methods can be expanded by using Auxiliary Deep Generative Models [63] or Normalizing Flows [78][55]. Doing so increases the expressibility of $p(z|y)$ - effectively re-expanding the search space over Z. However, these methods still suffer from the second drawback. But the authors of [2] have slightly reduced this problem by providing cases of networks where assumed marginals are almost correct.

GAN - like methods

Another class of Mutual Information Estimators/Controllers arise as a special case of f-GAN [72]. These rely on the following specified versions of Lemma 1 from reference [69]:

Lemma 1. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex lower semi-continuous function that is differentiable on the interior of its domain. Let $\phi^*(t)$ denote its convex conjugate. Let $\mathcal{F}$ denote a function class. Then*

$$\int \phi\left(\frac{d\left(\mathbb{P}_X \otimes \mathbb{P}_Z\right)}{d\mathbb{P}_{XZ}}\right) d\mathbb{P}_{XZ} \geq \sup_{f \in \mathcal{F}} \left[\int f d\left(\mathbb{P}_X \otimes \mathbb{P}_Z\right) - \int \phi^*(f) d\mathbb{P}_{XZ}\right] \tag{2.6}$$

with equality iff. $\phi'\left(\frac{d(\mathbb{P}_X \otimes \mathbb{P}_Z)}{d\mathbb{P}_{XZ}}\right) \in \mathcal{F}$. And

$$\int \phi\left(\frac{d\mathbb{P}_{XZ}}{d\left(\mathbb{P}_X \otimes \mathbb{P}_Z\right)}\right) d\left(\mathbb{P}_X \otimes \mathbb{P}_Z\right) \geq \sup_{f \in \mathcal{F}} \left[\int f d\mathbb{P}_{XZ} - \int \phi^*(f) d\left(\mathbb{P}_X \otimes \mathbb{P}_Z\right)\right] \tag{2.7}$$

with equality iff. $\phi'\left(\frac{d\mathbb{P}_{XZ}}{d(\mathbb{P}_X \otimes \mathbb{P}_Z)}\right) \in \mathcal{F}$.

The LHS of these bounds correspond directly to mutual information when $\phi(t)$ is equal to $-log(t)$ in Equation (2.6) and $\phi(t) = tlog(t)$ in Equation (2.7). In either of these two cases, if the corresponding optimization problem is consistent, then maximizing an empirical estimate of this functional then yields an estimator of mutual information. Reference [69] proved this consistency in the case of Equation (2.6) with $\phi(t) = -log(t)$.

But other ϕ can be used to estimate mutual information as well. Since the optimal solution to each objective function is $\phi'\left(\frac{d(\mathbb{P}_X \otimes \mathbb{P}_Z)}{d\mathbb{P}_{XZ}}\right)$ and $\phi'\left(\frac{d\mathbb{P}_{XZ}}{d(\mathbb{P}_X \otimes \mathbb{P}_Z)}\right)$ respectively, the empirically optimized function f^* will be an estimate of that derivative. In many cases, this derivative contains the log ratio $log\frac{d\mathbb{P}_{XZ}}{d(\mathbb{P}_X \otimes \mathbb{P}_Z)}$ directly. Then taking an empirical estimate of this log-ratio will yield an estimate of mutual information. For example, a regular GAN [39] can be used in this way as was done in reference [99] with the following objective function:

$$\inf_{r \in \mathcal{F}} \int -log\left(\sigma(r)\right) d\mathbb{P}_{XZ} + \int -log\left(1 - \sigma(r)\right) d\left(\mathbb{P}_X \otimes \mathbb{P}_Z\right) \tag{2.8}$$

Finally, we can substitute the convex conjugacy inequalities of Lemma 1 with the Donsker-Varadhan representation to achieve:

$$I(X;Z) \geq \sup_{f \in \mathcal{F}} \int f d\mathbb{P}_{XZ} - log \int e^f d\left(\mathbb{P}_X \otimes \mathbb{P}_Z\right) \tag{2.9}$$

optimizing this representation leads to biased gradients [9], but that problem is taken care of in the reference. Consistency of this estimator is proved in that paper as well. Further, while this representation has the same supremum as the convex conjugate representations, it is tighter for all in-optimal f [81].

2.1.3 Large Deviations Theory

The theory of large deviations concerns the exponential decay of probabilities of random variables far from their expectation. It can be viewed as a generalization of the central limit theorem from standard probability theory. Let $\{\mu_n\}$ be a family of measures on a Polish topological vector space $\mathcal{X}$. Then $\{\mu_n\}$ is said to satisfy the large deviations principal with rate function (or sometimes good-rate function) I if the following inequalities are satisfied:

$$-\inf_{x \in \Gamma^\circ} I(x) \leq \liminf_{n \to \infty} \frac{1}{n} log \mu_n(x \in \Gamma) \leq \limsup_{n \to \infty} \frac{1}{n} log \mu_n(x \in \Gamma) \leq -\inf_{x \in \bar{\Gamma}} I(x) \tag{2.10}$$

where the notation Γ° and $\bar{\Gamma}$ refer to the interior and closure of the set Γ respectively.

When a rate function is satisfied, we obtain the following approximate bounds on probabilities:

$$exp[-n \inf_{x \in \Gamma^\circ} I(x)] \leq \mu_n(x \in \Gamma) \leq exp[-n \inf_{x \in \bar{\Gamma}} I(x)] \tag{2.11}$$

which become tight as n approaches ∞. For $I : \mathcal{X} \to [0, \infty]$ to be a rate function, it is required that I be convex and lower-semicontinuous. Convexity is defined through satisfaction of the inequality $I(\theta x + (1-\theta)x') \leq \theta I(x) + (1-\theta) I(x')$ for all $\theta \in [0,1]$,

and all $(x, x') \in \mathcal{X} \times \mathcal{X}$. Lower semi-continuity is defined by satisfaction of the inequality $\underset{x \to x_0}{liminf}\ I(x) \geq I(x_0)$. I is said to be a good rate function if the inverse image of its level sets are compact. Large deviations principals can be derived for the case of $\mathcal{X}$ itself being a space of probability measures. This yields, for example, Sanovs theorem.

In several principal cases, the convex conjugate of the moment-generating function can be used as a good rate function:

$$\Lambda^*(x) = \underset{\lambda \in \mathcal{X}^*}{sup} \{\langle \lambda, x \rangle - log\ \mathbb{E}\left[e^{\langle \lambda, x \rangle}\right]\} \tag{2.12}$$

In the case of sequences of empirical probability measures on a space $\mathcal{X}$, this generalizes to:

$$\Lambda^*(\nu) = \underset{f}{sup} \{\mathbb{E}_\nu\left[f\right] - log\ \mathbb{E}_\mu\left[e^f\right]\} \tag{2.13}$$

where the supremum is taken over the space of either bounded continuous functions or bounded measureable functions on $\mathcal{X}$. This expression is equivalent to Kullback-Liebler divergence, and is known as the Donsker-Varadhan representation of KL divergence.

Denote the space of all Borel-probability measures on $\mathcal{X}$ as $M_1(\mathcal{X})$. If $M_1(\mathcal{X})$ is given the weak topology, and if we consider only compact, convex subsets of M_1, then the aforementioned large deviations bound holds non-asymptotically as well - that is, it will hold true for all finite n. All closed sets of $M_1(\mathcal{X})$ will be compact when $\mathcal{X}$ itself is compact, but we will often work with non-convex sets. In such cases, we can still find non-asymptotic bounds by covering our set with closed convex sets and then union bounding over those sets.

2.2 Notation and Assumptions

Capital letters denote random variables. Lower case letters describe instances of the corresponding random variable. Figure 2.1 depicts the classification model used in this dis-

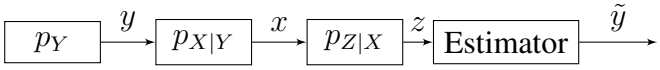

Figure 2.1: The classification model assumed in this dissertation.

sertation. A class variable y generates a feature vector x according to a fixed (unknown) distribution $\mathbb{P}_{X|Y}$. This feature vector is then fed through a learned distribution $\mathbb{P}_{Z|X}$, which acts as a lossy compressor of x. This should be thought of as the hidden layers of a neural network. z is then used to form an estimator of y, denoted $\tilde{y}$. We will drop the subscripts on probability distributions when the context is clear. The calligraphic symbols $\mathcal{X}$ and $\mathcal{Y}$ refer to the set of values that X and Y can take on. We assume that $\mathcal{X}$ is a Polish space such as $\mathbb{R}^d$ and that $\mathcal{Y}$ is a finite set with the discrete topology.

This model has three variables of interest, X, Y and Z which satisfy the Markov chain $Y - X - Z$. We denote the true model as $\mathbb{P}_{XYZ} = \mathbb{P}_X \mathbb{P}_{Z|X} \mathbb{P}_{Y|X}$ and consider the case of estimating the conditional probability distribution $\mathbb{P}_{Y|X}$. We denote this estimate as $\hat{\mathbb{P}}_{Y|X}$ and denote the estimated *full* model as $\hat{\mathbb{P}}_{XYZ} = \mathbb{P}_X \mathbb{P}_{Z|X} \hat{\mathbb{P}}_{Y|X}$. We will use the hat notation for all information theoretic quantities referring to the estimated model. For example:

$$\hat{I}(X;Y) := \mathbb{E}_{\hat{\mathbb{P}}_{XY}} \left[log \, \frac{d\hat{\mathbb{P}}_{XY}}{d\left(\mathbb{P}_X \otimes \hat{\mathbb{P}}_Y\right)} \right]$$

Finally, we assume that all distributions can be written as density functions such as $p_{XY}(x, y)$. We will occasionally drop the variable-specifying subscript when the context is clear. We will assume that the support of $p(x)$ is all of $\mathcal{X}$.

2.3 Background

2.3.1 The Information Bottleneck Principle

The use of the compressor $p_{Z|X}$ comes from the *Information Bottleneck Problem* [96] which attempts to find a variable Z that is **minimally sufficient** for the input pair of variables (X, Y). The minimal sufficiency of Z refers to the following two properties. First, X and Y must be conditionally independent given Z, or, put in a more enlightening way, $I(Z;Y) = I(X;Y)$. And second, for any other sufficient statistic T, $I(X;T) \geq I(X;Z)$. Intuitively, a minimally sufficient statistic is the most efficient description of X which retains all of the available information about the class variable Y. Further reasons that we wish to find a minimally sufficient statistic will become clear in the following sections.

2.3.2 Information and Generalization

We now focus on the reason for caring about the first aspect of finding a minimally sufficient statistic. That is, on finding a variable such that $I(Z;Y) = I(X;Y)$, or, in a more relaxed form, at least ensuring finding one such that $I(Z;Y)$ is relatively large. Pursuing this goal is backed by information theory as well as standard estimation theory. On the estimation theory side, this property just amounts to ensuring that Z be a sufficient statistic for X and Y. It thus has importance in finding optimal estimators, for example, through the Rao-Blackwell theorem [14]. On the information theoretic side, if $I(Y;Z) = H(Y)$, then having an instance z would completely determine the corresponding instance y, and so there exists an estimator of Y that takes Z as input and has zero probability of error. This notion can be expanded to $I(Y;Z) < H(Y)$ by Fano's inequality and its generalizations [20] [101]. Fano's inequality provides the following bound on estimation error for *any* estimator of Y:

$$h_2(P_e) + P_e log_2\left(|\mathcal{Y}| - 1\right) \geq H(Y) - I(Y;Z) \tag{2.14}$$

where P_e is the error rate of the estimator and h_2 denotes the binary entropy function $h_2(t) = -tlog_2(t) - (1-t)log_2(1-t)$. This inequality has a left hand side (LHS) that is strictly increasing in P_e for $P_e \leq \frac{1}{2}$. Thus the restriction of the LHS to $[0, \frac{1}{2}]$ is invertible, and since $H(Y)$ is fixed, we can say that P_e is lower bounded by a monotonically decreasing function of $I(Y;Z)$. In some cases we do achieve near equality in (2.14) - particularly when 1.) the estimator performs (nearly) equally well on each class and 2.) the estimator $Z \to \hat{Y}$ incurs relatively low levels of compression when compared to that which was incurred in the map $X \to Z$.

2.3.3 Information Losses

We now turn to the reason for caring about the second aspect of finding a minimally sufficient statistic - the minimality. This is where the role of our sampled data comes into play, and with it, the concept of information losses.

When we train on a finite sample of data, achieving the first aspect of a minimally sufficient statistic - the sufficiency - becomes difficult. This is because, no matter what representation we choose, we always have an information loss of the form:

$$I^{(1)}_{Loss} \triangleq |I(Y;Z) - \hat{I}(Y;Z)| \tag{2.15}$$

(The superscript (1) here is to distinguish between this form of information loss and another form which will appear later. We will call the current form *type one information losses*). In choosing our representation, we will only be able to control the latter term in this expression, as that term corresponds to the model we have estimated from our training data. Thus, if this loss is large, then, no matter what we do, we will have trouble in making $I(Y;Z)$ as large as possible. Throughout this dissertation, we will find that this term, $I^{(1)}_{Loss}$, depends on

$I(X;Z)$. In the old bounds (i.e. previous to this work), its dependence is exponential [89]:

$$I^{(1)}_{Loss} \leq \mathcal{O}\left(\sqrt{\frac{|\mathcal{Y}|}{2m}} 2^{I(X;Z)}\right) \tag{2.16}$$

where m is the number of training samples. And so we see that, at least in this form, keeping $I(X;Z)$ low is pertinent.

In this dissertation, we will find that the dependence on $I(X;Z)$ is relaxed to a linear one. Thus it may not always be so clear that we should minimize $I(X;Z)$. A perhaps more illuminating perspective can be found if we transfer instead to what we call *type two information losses*. These relate the best possible representation (in terms of achieving sufficiency) to the one that we would obtain by optimizing Z jointly with our estimated probability distribution. Before describing this new type of information loss, we will need to rigorously define the representations that we qualitatively described in the previous sentence.

Definition 1. *Let $\epsilon > 0$. We denote as $Z^*_\epsilon(I)$ and $\hat{Z}_\epsilon(I)$ any random variables that are at most ϵ-suboptimal for the following information bottleneck problems respectively:*

$$\sup_{p(z|x)} I(Y;Z)$$
$$\textit{subject to } I(X;Z) = I$$

$$\sup_{p(z|x)} \hat{I}(Y;Z)$$
$$\textit{subject to } I(X;Z) = I$$

We will then define type two information losses: $I^{(2)}_{Loss,\epsilon}(I) \triangleq I(Y;Z^*_\epsilon(I)) - I(Y;\hat{Z}_\epsilon(I))$, which are, in general, a function of $I \triangleq I(X;Z)$. Then, rearranging, we see that the quantity we care about, $I(Y;\hat{Z}_\epsilon(I))$, is given by $I(Y;Z^*_\epsilon(I)) - I^{(2)}_{Loss,\epsilon}(I)$, and so picking an $I(X;Z)$

that maximizes this expression is critical, though it may not always result in a direct minimization of $I(X;Z)$.

In any case, it is easy to convert bounds on type one information losses into corresponding bounds on type two information losses, as we will see in the next lemma.

Lemma 2. *Suppose that we have a bound of the form $I^{(1)}_{Loss} \leq K(\cdot)$, where $K(\cdot)$ can be any function of any number of arguments. Then:*

$$I^{(2)}_{Loss,\epsilon}(I) \leq 2K(\cdot) + \epsilon \tag{2.17}$$

2.3.4 Automatic Implementation via Neural Networks

There is evidence [92][3] that neural networks automatically solve the information bottleneck problem. The first set of evidence is experimental. Authors of [92] found that a wide range of neural networks undergo training in two phases. In the first phase, the neural networks memorized the inputs. This corresponded to an increase of $I(X;Z)$ and $I(Y;Z)$ simultaneously. During this phase, the average magnitude of back-propagated gradients surpassed the variance. In the second phase, this dynamic swapped and the variance surpassed the average. During this phase, $I(Y;Z)$ increased, but $I(X;Z)$ dropped - the neural networks were compressing the input to learn more about Y.

The second set of evidence is theoretical. The authors of [3] show that $I(X;Z)$ is tightly related to the information between the weights and the data $I(W;\mathcal{D}^l)$. This relationship holds with only a few assumptions on the corresponding neural network. They then shown that $I(W;\mathcal{D}^l)$ is small when the network converges to a *wide* local minimum of the cross entropy loss function. Finally, they argue that stochastic gradient descent tends to converge to such minima.

Some more recent experimental evidence [83] counters these two arguments. This new evidence shows that some networks can achieve high $I(Y;Z)$ without compression. Thus some networks can significantly outperform the lower bound of inequality (2.16). This dissertation presents new lower bounds which are much tighter and less sensitive to $I(X;Z)$ than (2.16). These bounds - while useful on their own right- help to explain this counter evidence.

2.3.5 Training Data Selection Related Works

The subject of training data selection is extensive. We will consider a coarse division of the field. On one side of this divide is batch mode learning, which selects data all at once. Methods on the batch mode side include the collection of literature on sensor placement [40, 58], facility location based methods [85], and transductive experimental design [108]. On the other side of the divide is active learning, which selects new data in iteration by training a new classifier on the currently selected data. Most methods of this type follow from a powerful idea: label the data points that our current classifier is most uncertain of [30, 36, 37, 46, 52, 59, 66, 88]. Much can be found in comprehensive texts [86, 98] and literature surveys [87]. While most work in the field of training data selection falls on the active learning side, our framework is applicable to both parts of the division, and we will give a slight focus to the batch mode side.

The field is ripe with active learning algorithms that are highly justified within the classical/PAC statistical learning theory [15]. Beginning with the CAS algorithm [19], and being subsequently improved upon in terms of applicability [7, 12, 13, 23], this branch of work rigorously derives algorithms which obtain label complexities, for seperable data, of $O(\theta d \, log \, \frac{1}{\epsilon})$ where d is the VC dimension of the hypothesis space, ϵ is the desired classification error, and θ is a useful quantity called the disagreement coefficient of the dataset/hypothesis space pair [42]. This is an exponential improvement over the label

complexity required of random labelling, which needs $O(\frac{d}{\epsilon})$ labels for the same error rate under the same classical learning theory.

Some early work in the above path even uses information theoretic notions [35, 38]. Specifically, they maintain a probability distribution over the hypothesis space, and data is selected such that the entropy of that distribution is minimized when conditioned on the event $\{h : h \text{ is consistent with the labelled data}\}$. Unfortunately, this notion of information is not placed on the class/representation variables themselves, and so they cannot use Fano's inequality in assessing their complexity - instead, they also rely on classical learning theory, obtaining complexities again on the order of $O(d\ log\ \frac{1}{\epsilon})$ while having the additional complication of needing to maintain and sample from a sequence of posterior distributions on the hypothesis space.

While the above analyses are fantastic for machine learning algorithms which conform to classical learning theory, there is a problem with adapting them to deep learning methods. Classical learning theory does not appear to predict the empirical effectiveness of deep learning methods. For example, while the size of a network grows, d increases quite quickly, but the label complexity of the learner drops in experiment, even in the randomly selected case. Thus we turn to a promising emerging field of learning theory which relates deep learning to information theory upon which we build our framework [3, 33, 89, 92, 97]. And of course, many of those active learning methods derived from classical learning theory may be analyzed with this new framework, perhaps giving more satisfying label complexities when applied to deep learning. To our knowledge, there is no previous work in data selection theory which employs this more modern theory of deep learning.

2.3.6 Phase Identification Related Works

Much phase identification work is based on physical approaches [6, 16, 17, 103]. References [17] and [6] develop phase identification systems based on high resolution timing mea-

surements communicated between the base station and the feeder transformer secondaries and/or individual electricity consumers. This system is highly accurate and even yields the voltage phasors of the secondaries themselves instead of just the phase names of the wires connected to them. However, the system is quite sophisticated. Deploying such a system for each feeder across the many distribution circuits overseen by a utility is very expensive.

Reference [103] describes a phase identification technique using micro-synchrophasors. The overhead cost of this method is less than the that of reference [17] since micro-synchrophasors are mobile - hence only a few devices are required. However, this reduced overhead cost results in increased labor and time costs, as the micro-synchrophasors must be reinstalled several times throughout the distribution network. Reference [16] patents a method for phase identification through signal injection. A signal generator is placed at the base substation and a unique signal is created for each phase. These signals are detected by a signal discriminator at each customer location. By matching the signals, the phase connectivities can be accurately reproduced. Like reference [17], this method is very intensive in labor and time, but relatively cheap in overhead. These physical methods are the approaches that are typically taken during the field testing projects employed in practice. Furthermore, they act as the base upon which training labels should be acquired.

The amount of literature related to solving of phase identification problem with data-driven methods is more limited. Of what does exist, most is unsupervised. This unsupervised branch can be split further into 'model-agnostic' and 'model-based' sub-branches.

In the model-based sub-branch, reference [25] compares simulated power-flow solutions for a given phase configurations to real data. This method is accurate, but requires a correct system model including everything except phase connectivity's - e.g. line parameters, network topology. References [1] and [5] group customers by phase such that the total sum of power injections on each phase matches that of the substation or distribution transformers up to some error. These methods are somewhat non-robust to missing or erroneous data

since these will lead to power mismatches between the measured load and the measured supply. Furthermore, phase configuration is often assumed to already exist in methods of network topology estimation and parameter estimation, so we typically like to think of phase identification as the first step in a long sequence of modelling techniques. Assuming realistic models as an input to the phase identification problem is somewhat unrealistic at this time.

In the model-agnostic sub-branch, one of the most popular techniques is to correlate voltage time series data at a household to the voltage time series data captured with other households [62, 74, 75, 84, 107]. These works showcase the power of using voltage data as a primary predictor of phase type - an idea which we will study in depth in this dissertation. However, while the statement 'customer A and customer B are on the same phase implies customer A and customer B have correlated voltages' is mostly accurate, these methods struggle in two ways. First, the converse of this statement is not always accurate. That is, customers can be on different phases and still have correlated voltages. Second, knowing that two customers are on the same phase doesn't actually tell us what that phase *is*. This second issue can be resolved by either physical inspection, which are time and labor intensive, or by correlating the relevant voltages to those of the substation [91]. But substation voltages are usually only measured either line to line or line to neutral - not both simultaneously, and so this limits the scope of circuits that we are capable of classifying to a small set of edge cases.

Finally, some unsupervised clustering techniques [70, 102] and frequency domain feature extraction techniques [107] have been tested on this problem. Supervised methods are limited to off-the-shelf algorithms that don't take any consideration into the properties of power distribution networks. To the author's knowledge, the method presented in this dissertation is the first of its type.

2.3.7 Battery Storage Literature Review

While much research has been done on the value of BESS systems directly linked to renewable sources [71] [60], for energy arbitrage [4], and for primary frequency regulation [73], a large amount of the existing literature has underestimated potential profits by failing to optimize these actions simultaneously. By correcting these issues, [109] and [104] obtained a significantly more optimal valuation of BESS. However, this result was obtained by considering only the effect of time on the battery's deterioration. But actions themselves cause significant wear and tear via cycle degradation. Optimizing over just the space of actions will directly cause the battery to under perform later in its life or even fail prematurely. Thus the result of that analysis is likely an overestimate.

Only a few attempts have been made to couple a BESS ancillary service optimization problem with degradation. However, there is research in degradation models suitable to other optimization use cases.

Reference [8] introduces one of the lowest level degradation models suitable for application. The model considers a current driven differential equation representing build up of resistance at the battery anode. This is coupled with a low level battery discharge model to find the driving current. The model was used to find optimal charging schemes for electric vehicles. Since it is such a low level model, it is difficult to use in more complicated optimization schemes. It also only considers resistive build up, and thus does not consider capacity loss. Nonetheless, if the optimization use case is based directly on charging and discharging profiles with little uncertainty, then this is a good model with excellent theoretical justification.

Most models describe degradation in terms of cycling parameters. Reference [29] provides a comprehensive, test driven analysis of battery degradation based on these parameters. It includes analysis in both capacity loss and resistive build up. The paper does not provide

an analytical model for degradation, but does discuss useful insights for the effects of each cycle parameter of degradation. These insights can be used to develop semi-analytical fitting models for degradation.

Reference [90] presents a simple parameter based model for predicting battery lifetime under a uniform cycling scheme. For nonuniform cycling, a distribution of current rates is assumed. The lifetime model can be easily converted to a degradation model. It is used in the paper to develop a power control strategy for a hybrid battery/utra-capacitor storage system in electric vehicles. It is a suitable model when cycling parameter distributions can be assumed. However, it is unclear if this model will remain useful in optimization problems where the decisions change the profile of cycles.

Reference [47] develops another cycle based degradation model for the optimization of charging profiles in electric vehicles. The optimization uses time varying electricity costs and estimated degradation costs. It finds various charging optimal charging profiles based on the form of these costs.

Reference [76] develops a degradation model based directly on state of charge for use in an economic optimization of a hybrid battery/ photo-voltaic system. Though the functional form of state of charge degradation is complicated in this model, it is useful because many applications can use the battery's state of charge directly as a decision variable. The model's main disadvantage is that it does not have any direct consideration of depth of discharge. The optimization procedure took care of depth of discharge loss by putting upper and lower bounds on the state of charge.

A similar optimization problem was considered in [68]. In this paper, a semi-analytical degradation model, based on the properties found in [29], is developed. Through rainflow counting on real data, this paper finds a distribution of cycle parameters and uses this to find optimal charging profiles in hybrid battery/photo-voltaic systems. The model is also used to determine the best of three possible charging profiles for mobile phone longevity.

Other cycle parameter degradation models include reference [22], which optimizes a hybrid battery/HVAC system scheduling with battery degradation included, and reference [111] which considers battery degradation in an economic optimization of battery integration in a standalone microgrid.

Some work has been done in coupling ancillary services with degradation. Most of these focus on battery control while providing ancillary services rather than on deciding how much ancillary service to provide at a given time. For example, reference [34] coupled degradation to a control strategy for peak shaving. In particular, reference [64] formulated frequency regulation as a nonlinear tracking problem and included a degradation model as a state variable which is to be driven to zero during the tracking. We believe these works to be applicable in conjunction with the work presented here by using the regulation decisions from our scheme with the control strategies of those.

Reference [44] is the closest work to our own. It derives a simplified (though still nonlinear) degradation model that avoids rainflow counting and embeds it into an ancillary service optimization problem. The paper found good results and is a great exploration into coupling degradation considerations with ancilliary service scheduling. However, the results are found by optimizing over just one representative day and multiplying the daily profit by the lifetime (in days) that the representative day's operation would yield. The critical limitation to this approach is that it assumes lifetime can be accurately calculated from one day of battery operation. Realistic operation of a battery will vary day to day, and so a degradation model that considers the operations of each day is necessary.

2.3.8 Battery Modelling

Several circuit equivalent models for rechargeable batteries exist in literature. Of particular interest is the Dual Polarization (DP) Thevenin Equivalent model [45] shown in Figure 1.1. The two capacitors of the DP Model represent two independent types of capacity. The First

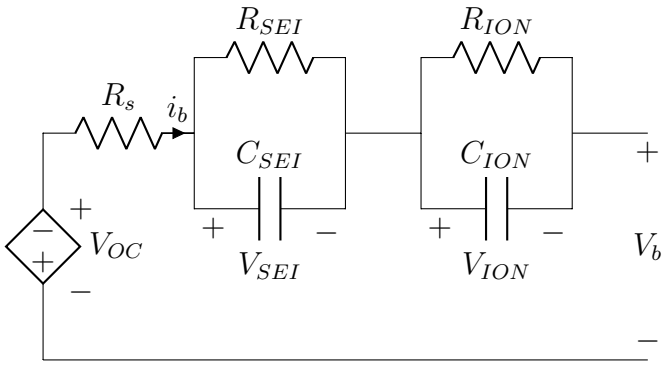

Figure 2.2: The DP Thevenin Equivalent Model of a battery

is interpreted as the Surface Interface Exchange (SEI) Layer capacity and the second is interpreted as the ionic capacity (for lead-acid batteries). This double-capacity model works well with empirical degradation models. In such models, maximum battery capacity takes the form of the sum of two independently decaying exponentials. This corresponds to the independent degradation of each capacitor subcircuit.

V_{OC} is a nonlinear function of the state of charge [110]

$$SoC = \frac{1}{Q_{max}}(Q_{SEI} + Q_{ION}) \tag{2.18}$$

where

$$Q_{SEI} = C_{SEI}V_{SEI} \tag{2.19}$$

$$Q_{ION} = C_{ION}V_{ION} \tag{2.20}$$

and Q_{max} is a known maximum battery capacity. Empirical data of the open circuit voltage with respect to SoC yield curves of the form shown in Figure 1.2. This function has nearly affine behavior in the interval $(20,\ 80)$. Thus, for this interval, we can write $V_{OC}(SoC) = a_0 + a_1 SoC$. The DP Equivalent circuit leads to a diagonalized state space model. We take as state variables the following four quantities taken from this circuit's

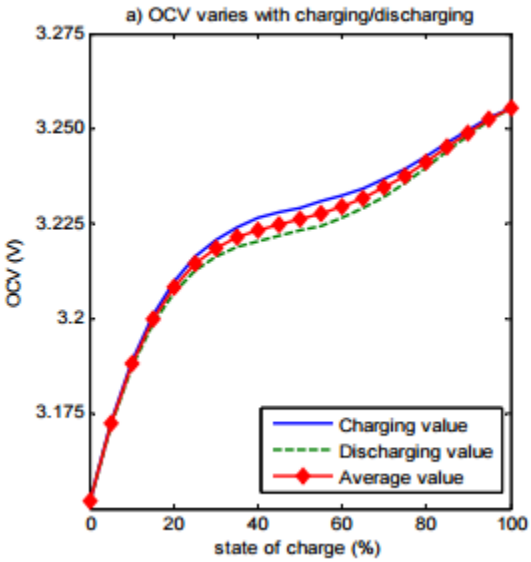

Figure 2.3: Open Circuit Votlage with respect to State of Charge. Figure from [110].

measurable values.

$$\zeta_{SEI} = Q_{SEI} - \frac{1}{2}\gamma_{SEI} \tag{2.21}$$

$$\zeta_{ION} = Q_{ION} - \frac{1}{2}\gamma_{ION} \tag{2.22}$$

$$\gamma_{SEI} = (\frac{C_{SEI}}{C_{SEI} + C_{ION}})Q_{max} \tag{2.23}$$

$$\gamma_{ION} = (\frac{C_{ION}}{C_{SEI} + C_{ION}})Q_{max} \tag{2.24}$$

In this representation, the first two variables are zero if and only if $SoC = \frac{1}{2}Q_{max}$.

The state space representation is then:

$$\zeta = \mathcal{A}\zeta + \mathcal{B}i_b(t) \tag{2.25}$$

$$(V_b - a_0) = \mathcal{C}\zeta + \mathcal{D}i_b(t) \tag{2.26}$$

where

$$\zeta = \begin{bmatrix} \zeta_{SEI} & \zeta_{ION} & \gamma_{SEI} & \gamma_{ION} \end{bmatrix}^T \tag{2.27}$$

$$\mathcal{A} = \begin{bmatrix} -\frac{1}{C_{SEI}R_{SEI}} & 0 & -\frac{1}{2C_{SEI}R_{SEI}} & 0 \\ 0 & -\frac{1}{C_{ION}R_{ION}} & 0 & -\frac{1}{2C_{ION}R_{ION}} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.28}$$

$$\mathcal{B} = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}^T \tag{2.29}$$

$$\mathcal{C} = \begin{bmatrix} (\frac{a_1}{Q_{max}} - \frac{1}{C_{SEI}}) & (\frac{a_1}{Q_{max}} - \frac{1}{C_{ION}}) \end{bmatrix} \tag{2.30}$$

$$\mathcal{D} = -R_s \tag{2.31}$$

2.3.9 Degradation Modeling

Battery degradation occurs as a (non-analytical) function of the state of charge time-series $SoC(k)$. It is calculated in two steps. First, the time-series is converted to a set of cycles through the Rainflow Counting Algorithm (RCA). The cycles are then converted into degradation factors which are then used to calculate a new maximum battery capacity.

Rainflow Counting

The RCA is used to detect repetition in an aperiodic time series and compute a list of cycles, nonuniform in amplitude and duration, that are embedded within each other. Each cycle in this list is represented as an ordered set of parameter values. For battery degradation, the useful parameters are the Depth of Discharge (DoD), defined as the amplitude of a cycle normalized to the maximum battery capacity, the mean SoC, defined as the average of the cycle with time taken into account, and the current rate (CR), defined as the DoD divided by

the cycle duration. Typically, the first step of the RCA reduces the input time series to a set of peak and valley values. This removes all knowledge of time, and so duration can not be calculated with this set up. To fix this, we have augmented the first step so that it keeps the intra-hour time indices at which the peaks and valleys occurred.

Capacity Calculation

The new maximum capacity is obtained from the outputs of the RCA as follows. First, for each cycle (indexed by i), the following three semi-empirical functions are calculated

$$f_{DoD}(DoD_i) = (k_{DoD,1} DoD_i^{-k_{DoD,2}} - k_{DoD,3})^{-1} \tag{2.32}$$

$$f_{SoC}(SoC_i) = e^{k_{SoC}(SoC_i - SoC_{ref})} \tag{2.33}$$

$$f_{CR}(CR_i) = e^{k_{CR}(CR_i - CR_{ref})} \tag{2.34}$$

Where DoD_i, SoC_i, and CR_i are the DoD, SoC and CR of the i^{th} cycle. The forms of these functions are derived from theoretical considerations. Parameter values for Lithium-Ion (Li-Ion) batteries were found experimentally in [106]. Table 2.1 repeats these values for convenience.

A degradation rate deg is then calculated from these functions according to (2.35).

$$deg = \sum_{i=1}^{L} f_{DoD}(DoD_i) f_{SoC}(SoC_i) f_{CR}(CR_i) + k_t T \tag{2.35}$$

where L is the number of cycles returned from the RCA, k_t is the rate at which the battery ages independently from operation, and T is the length of the interval upon which the SoC time series is defined. We will call deg the *degradation function*, viewing it as a function of cycles and time.

Table 2.1: Degradation function parameters for a Li-Ion Battery.

Function Parameters		
Stressor	Parameter	Value
DoD	$k_{DoD,1}$	8.95×10^4
	$k_{DoD,2}$	4.86×10^{-1}
	$k_{DoD,3}$	7.28×10^4
SoC	k_{SoC}	1.04
	SoC_{ref}	0.50
CR	k_{CR}	2.63×10^{-1}
	CR_{ref}	1.00
Time	k_t	$1.49 \times 10^{-6} \frac{1}{hr}$
r_1		5.75×10^{-2}
r_2		121

Finally, the new capacity is calculated from the following double exponential model.

$$E_{max}^{(n+1)} = r_1 e^{-r_2 \sum_{\eta=1}^{n} deg_\eta} + (1 - r_1) e^{\sum_{\eta=1}^{n} deg_\eta} \tag{2.36}$$

The first exponential represents a quick degradation from the build up of the Solid Electrolyte Interphase (SEI) layer. The second represents a slower degradation from ion loss. Values for r_1 and r_2 are provided in Table 2.1. The sum term in the exponentials takes in all previously calculated degradations.

CHAPTER 3

BOUNDS ON INFORMATION LOSSES

3.1 Product Form Decomposition - Setup

Our first major step is a decomposition of information losses into a product of two terms, one being $I(X;Z)$, and the other being a term related to a statistical distance between $\mathbb{P}$ and $\hat{\mathbb{P}}$. The proof of this decomposition takes some setting up. The setup is performed by generalizing the well studied maximal coupling [82] from statistics to our purposes. We will call our generalization the *conditional maximal coupling*, and will begin its construction by quickly reviewing couplings in general [48].

Definition 2 (Coupling)**.** *Given two probability models $\mathbb{P}_{\tilde{S}}$ and $\mathbb{Q}_S$ on a list of variables S, a **coupling** of these models is a pair of random variables $(\tilde{S}, \hat{S})$ with joint distribution $\gamma_{\tilde{S},\hat{S}}$ such that the marginal distributions satisfy $\gamma_{\tilde{S}} = \mathbb{P}_{\tilde{S}}$ and $\gamma_{\hat{S}} = \mathbb{Q}_S$.*

Couplings, as used in this dissertation, are convenient because they allow us to manipulate integral quantities relating the true and estimated models. For example,

$$\int f(p(s))d\mathbb{P}_S - \int f(q(s))d\mathbb{Q}_S = \int \left(f(p(\tilde{s})) - f(q(\hat{s}))\right) d\gamma \tag{3.1}$$

We will be dealing with a specific coupling which is derivative of the well studied maximal coupling [82]. The emphasis is that, in translating from $\mathbb{P}_{XYZ}$ to $\hat{\mathbb{P}}_{XYZ}$, only $\mathbb{P}_{Y|X}$ changes. Thus we focus on coupling $\mathbb{P}_{Y|X}$ to $\hat{\mathbb{P}}_{Y|X}$ while leaving the rest of the model unchanged.

Construction 1 (Conditional Maximal Coupling)**.** *We set our coupling to consist of two triples of random variables. The first is denoted as* $(\tilde{X}, \tilde{Y}, \tilde{Z})$ *and the second is denoted as* $(\hat{X}, \hat{Y}, \hat{Z})$*). These are defined as follows. First, define the function* $m_l : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ *through*

$$m_l(a, b) := \textit{min}\{p_{Y|X}(b|a), \hat{p}_{Y|X}(b|a)\} \tag{3.2}$$

Next, define a real number ρ *as*

$$\rho := \int \left(\sum_y m_l(x, y) \right) d\mathbb{P}_X \tag{3.3}$$

and define J *as a Bernoulli random variable with success probability* ρ*. Then define variables* $U = (U_1, U_2), V = (V_1, V_2)$ *and* $W = (W_1, W_2)$ *through*

$$p_{U_1,U_2}(u_1, u_2) := \frac{p_X(u_1) m_l(u_1, u_2)}{\rho} \tag{3.4}$$

$$p_{V_1,V_2}(v_1, v_2) := \frac{p_X(v_1) p_{Y|X}(v_2|v_1) - p_X(v_1) m_l(v_1, v_2)}{1 - \rho} \tag{3.5}$$

$$p_{W_1,W_2}(w_1, w_2) := \frac{p_X(w_1) \hat{p}_{Y|X}(w_2|w_1) - p_X(w_1) m_l(w_1, w_2)}{1 - \rho} \tag{3.6}$$

Next define $(\tilde{X}, \tilde{Y}, \hat{X}, \hat{Y})$ *as functions of the above random variables as follows:*

$$\begin{cases} (\tilde{X}, \tilde{Y}) = (\hat{X}, \hat{Y}) = (U_1, U_2) & \textit{if } J = 1 \\ (\tilde{X}, \tilde{Y}) = (V_1, V_2),\ (\hat{X}, \hat{Y}) = (W_1, W_2), & \textit{if } J = 0 \end{cases} \tag{3.7}$$

Finally, we define $\tilde{Z}$ and $\hat{Z}$ through:

$$\gamma_{\hat{Z}|\hat{X}} = \gamma_{\tilde{Z}|\tilde{X}} = p_{Z|X} \tag{3.8}$$

Lemma 3. *Construction 1 yields a valid coupling.*

Proof. We first check that the defined variables J, U, V and W have valid distributions. For J to be valid, we need only check that $\rho < 1$. Indeed by replacing the min operation in $m_l(x, y)$ with $p_{Y|X}(y|x)$, we have that $\rho = \int \left(\sum_y m_l(x, y)\right) d\mathbb{P}_X \leq \int d\mathbb{P}_{XY} = 1$. The variable U is similarly valid since $\int d\mathbb{P}_U = \frac{1}{\rho} \int \left(\sum_y m_l(u_1, u_2)\right) d\mathbb{P}_X = \frac{\rho}{\rho} = 1$. The variables V and W follow similarly since $\int d\mathbb{P}_V = \frac{1}{1-\rho} \left(\int d\mathbb{P}_{XY} - \rho\right) = 1$ and $\int d\mathbb{P}_W = \frac{1}{1-\rho} \left(\int d\hat{\mathbb{P}}_{XY} - \rho\right) = 1$.

We then need to show that the marginals of the coupling satisfy $\gamma_{\tilde{X},\tilde{Y},\tilde{Z}} = \mathbb{P}_{XYZ}$ and $\gamma_{\hat{X},\hat{Y},\hat{Z}} = \hat{\mathbb{P}}_{XYZ}$. To begin, we first show that $\gamma_{\tilde{X},\tilde{Y}}(x, y) = p_{X,Y}(x, y)$ and that $\gamma_{\hat{X},\hat{Y}}(x, y) = \hat{p}_{X,Y}(x, y)$ as follows:

$$\gamma_{\tilde{X},\tilde{Y}}(x, y) = \rho \frac{p(x)m_l(x, y)}{\rho} + (1 - \rho) \frac{p(x)p(y|x) - p(x)m_l(x, y)}{1 - \rho} = p(x, y) \tag{3.9}$$

$$\gamma_{\hat{X},\hat{Y}}(x, y) = \rho \frac{p(x)m_l(x, y)}{\rho} + (1 - \rho) \frac{p(x)\hat{p}(y|x) - p(x)m_l(x, y)}{1 - \rho} = \hat{p}(x, y). \tag{3.10}$$

Finally, since we defined $\tilde{Z}$ and $\hat{Z}$ through the distributions by having both $\gamma_{\tilde{Z}|\tilde{X}}(z|x)$ and $\gamma_{\hat{Z}|\hat{X}}(z|x)$ equal to $p(z|x)$, we have that:

$$\gamma_{\tilde{X},\tilde{Y},\tilde{Z}}(x, y, z) = \gamma_{\tilde{X},\tilde{Y}}(x, y)\gamma_{\tilde{Z}|\tilde{X}}(z|x) = p(x, y)p(z|x) = p(x, y, z) \tag{3.11}$$

and

$$\gamma_{\tilde{X},\tilde{Y},\tilde{Z}}(x,y,z) = \gamma_{\tilde{X},\tilde{Y}}(x,y)\gamma_{\tilde{Z}|\tilde{X}}(z|x) = p(x,y)p(z|x) = p(x,y,z) \tag{3.12}$$

$$\gamma_{\hat{X},\hat{Y},\hat{Z}}(x,y,z) = \gamma_{\hat{X},\hat{Y}}(x,y)\gamma_{\hat{Z}|\hat{X}}(z|x) = \hat{p}(x,y)\hat{p}(z|x) = \hat{p}(x,y,z) \tag{3.13}$$

completing the proof. □

Lemma 4 (Coupling-Total Variation)**.** *The definitions of Construction 1 satisfy the following relationship:*

$$1 - \rho = \gamma\left(\tilde{Y} \neq \hat{Y}\right) = \mathbb{E}_{\mathbb{P}_X}\left[\frac{1}{2}\sum_y |p(y|x) - \hat{p}(y|x)|\right] \tag{3.14}$$

Proof. To prove the first equality, define the following subsets of $\mathcal{Y}$.

$$A(x) := \{y : p(y|x) \leq \hat{p}(y|x)\} \tag{3.15}$$

Then for any coupling of these two models:

$$\begin{aligned}
\mathbb{P}(\tilde{Y} = \hat{Y}|\tilde{X} = \hat{X} = x) &= \mathbb{P}(\tilde{Y} = \hat{Y}, \tilde{Y} \in A(x)|\tilde{X} = \hat{X} = x) \\
&\quad + \mathbb{P}(\tilde{Y} = \hat{Y}, \tilde{Y} \in A^c(x)|\tilde{X} = \hat{X} = x) \\
&\leq \mathbb{P}(\tilde{Y} \in A(x)|\tilde{X} = \hat{X} = x) \\
&\quad + \mathbb{P}(\hat{Y} \in A^c(x)|\tilde{X} = \hat{X} = x) \\
&= \sum_{y \in A(x)} p(y|x) + \sum_{y \in A^c(x)} \hat{p}(y|x) \\
&= \sum_y \min\{p(y|x), \hat{p}(y|x)\} = \sum_y m_l(x,y)
\end{aligned} \tag{3.16}$$

It follows that $\mathbb{P}(\tilde{Y} = \hat{Y}) = \int_X \mathbb{P}(\tilde{Y} = \hat{Y} | \tilde{X} = \hat{X} = x) d\mathbb{P}_X \leq \rho$. But we also have for this particular coupling that $\mathbb{P}(\tilde{Y} = \hat{Y}) \geq P_J(1) = \rho$. Thus we must have equality.

To prove the second equality, we will use the fact that $\min\{a, b\} = \frac{a+b-|a-b|}{2}$. Then we have:

$$\begin{aligned} \sum_y m(x, y) &= \frac{1}{2} \sum_y \left(p(y|x) + \hat{p}(y|x) - |p(y|x) - \hat{p}(y|x)| \right) \\ &= 1 - \frac{1}{2} \sum_y |p(y|x) - \hat{p}(y|x)| \end{aligned} \tag{3.17}$$

Thus $\rho = 1 - \mathbb{E}_{\mathbb{P}_X} \left[\frac{1}{2} \sum_y |p(y|x) - \hat{p}(y|x)| \right]$, completing the proof. □

3.1.1 Product Form Decomposition - Theorem and Proof

Motivated by Lemma 4, we will denote $1 - \rho$ as $\bar{\delta}(\hat{\mathbb{P}})$. This notation emphasizes its role as an average total variation distance. This finishes our setup for the decomposition, which we will now move on to prove.

Theorem 1.

$$\left| I(Y; Z) - \hat{I}(Y; Z) \right| \leq \bar{\delta}(\hat{\mathbb{P}}) I(X; Z) + h_2 \left(\bar{\delta}(\hat{\mathbb{P}}) \right) \tag{3.18}$$

We will use several Markov chains in this proof. All of them follow from the Bayesian network describing the generative process of all relevant random variables, which is depicted in figure 3.1. Each Markov chain that we use comes from the fact that the X variables d-separate the Z variables from the rest of the network. First, via coupling, we have that

Proof.

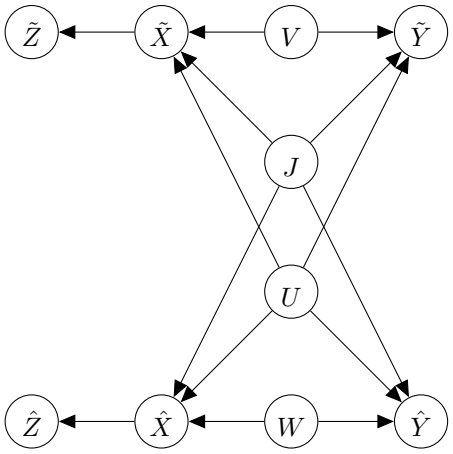

Figure 3.1: Bayesian Network describing the relationships between random variables in the proof of Theorem 1.

$\left|I(Y;Z) - \hat{I}(Y;Z)\right| = \left|I(\tilde{Y};\tilde{Z}) - I(\hat{Y};\hat{Z})\right|$. We decompose these terms as follows:

$$I(\tilde{Y};\tilde{Z}) = I(\tilde{Y};\tilde{Z}|\tilde{X}) + I(\tilde{X};\tilde{Z}) - I(\tilde{X};\tilde{Z}|\tilde{Y}) \tag{3.19}$$

$$I(\hat{Y};\hat{Z}) = I(\hat{Y};\hat{Z}|\hat{X}) + I(\hat{X};\hat{Z}) - I(\hat{X};\hat{Z}|\hat{Y}) \tag{3.20}$$

But, due to the Markov chains $\tilde{Z} - \tilde{X} - \tilde{Y}$ and $\hat{Z} - \hat{X} - \hat{Y}$, we have that $I(\tilde{Y};\tilde{Z}|\tilde{X}) = I(\hat{Y};\hat{Z}|\hat{X}) = 0$. Furthermore, $I(\tilde{X};\tilde{Z}) = I(\hat{X};\hat{Z}) = I(X;Z)$. It follows that:

$$\left|I(\tilde{Y};\tilde{Z}) - I(\hat{Y};\hat{Z})\right| = \left|I(\hat{X};\hat{Z}|\hat{Y}) - I(\tilde{X};\tilde{Z}|\tilde{Y})\right| \tag{3.21}$$

We can further decompose each of these terms as:

$$\begin{aligned} I(\hat{X};\hat{Z}|\hat{Y}) &= I(\hat{Z};\hat{X}|J,\hat{Y}) + I(\hat{Z};J|\hat{Y}) - I(\hat{Z};J|\hat{X},\hat{Y}) \\ I(\tilde{X};\tilde{Z}|\tilde{Y}) &= I(\tilde{Z};\tilde{X}|J,\tilde{Y}) + I(\tilde{Z};J|\tilde{Y}) - I(\tilde{Z};J|\tilde{X},\tilde{Y}) \end{aligned} \tag{3.22}$$

But we have from the Markov chains $\hat{Z} - \hat{X} - J$ and $\tilde{Z} - \tilde{X} - J$ that $I(\hat{Z}; J|\hat{X}, \hat{Y}) = I(\tilde{Z}; J|\tilde{X}, \tilde{Y}) = 0$. We can then break down $I(\hat{Z}; \hat{X}|J, \hat{Y})$ and $I(\tilde{Z}; \tilde{X}|J, \tilde{Y})$ as:

$$\begin{aligned} I(\hat{Z}; \hat{X}|J, \hat{Y}) &= \rho I(\hat{Z}; \hat{X}|J = 1, \hat{Y}) + (1 - \rho) I(\hat{Z}; \hat{X}|J = 0, \hat{Y}) \\ &= \rho I(\hat{Z}; U_1|U_2) + \bar{\delta}(\hat{\mathbb{P}}) I(\hat{Z}; W_1|W_2) \end{aligned} \tag{3.23}$$

and

$$I(\tilde{Z}; \tilde{X}|J, \tilde{Y}) = \rho I(\tilde{Z}; U_1|U_2) + \bar{\delta}(\hat{\mathbb{P}}) I(\tilde{Z}; V_1|V_2) \tag{3.24}$$

But when $\tilde{X} = \hat{X} = U_1$, $I(\hat{Z}; U_1|U_2) = I(\tilde{Z}; U_1|U_2)$. Thus we can decompose the term $\left| I(Y; Z) - \hat{I}(Y; Z) \right|$ to:

$$\begin{aligned} &\left| \bar{\delta}(\hat{\mathbb{P}}) \left(I(\hat{Z}; W_1|W_2) - I(\tilde{Z}; V_1|V_2) \right) + I(\hat{Z}; J|\hat{Y}) - I(\tilde{Z}; J|\tilde{Y}) \right| \\ \leq &\bar{\delta}(\hat{\mathbb{P}}) \left| I(\hat{Z}; W_1|W_2) - I(\tilde{Z}; V_1|V_2) \right| + \left| I(\hat{Z}; J|\hat{Y}) - I(\tilde{Z}; J|\tilde{Y}) \right| \end{aligned} \tag{3.25}$$

Now, from the Markov chains $\hat{Z} - \hat{X} - W_1$, $\hat{Z} - \hat{X} - W_2$, $\tilde{Z} - \tilde{X} - V_1$, and $\tilde{Z} - \tilde{X} - V_2$, we have (via applications of the data processing inequality and its corollaries [20]) that:

$$I(\hat{Z}; W_1|W_2) \leq I(\hat{Z}; \hat{X}|W_2) \leq I(\hat{Z}; \hat{X}) = I(X; Z) \tag{3.26}$$

$$I(\tilde{Z}; V_1|V_2) \leq I(\tilde{Z}; \tilde{X}|V_2) \leq I(\tilde{Z}; \tilde{X}) = I(X; Z) \tag{3.27}$$

Further, $I(\hat{Z}; J|\hat{Y}) \leq H(J)$ and $I(\tilde{Z}; J|\tilde{Y}) \leq H(J)$. It follows that:

$$\left| I(Z; W_1|W_2) - I(\tilde{Z}; V_1|V_2) \right| \leq I(X; Z) \tag{3.28}$$

$$\left| I(\hat{Z}; J|\hat{Y}) - I(\tilde{Z}; J|\tilde{Y}) \right| \leq H(J) = h_2 \left(\bar{\delta}(\hat{\mathbb{P}}) \right) \tag{3.29}$$

And so, in total, we have:

$$\left|I(Y;Z) - \hat{I}(Y;Z)\right| \leq \bar{\delta}(\hat{\mathbb{P}})I(X;Z) + h_2\left(\bar{\delta}(\hat{\mathbb{P}})\right) \tag{3.30}$$

completing the proof. □

A potentially useful special case of this bound occurs when we set $Z = X$:

Corollary 1. *If X is discrete,*

$$\left|I(X;Y) - \hat{I}(X;Y)\right| \leq \bar{\delta}(\hat{\mathbb{P}})H(X) + h_2(\bar{\delta}(\hat{\mathbb{P}})) \tag{3.31}$$

3.1.2 Understanding $\bar{\delta}(\hat{\mathbb{P}})$

The above relationships *looks* linear on $I(X;Z)$. However, $\hat{p}(y|x)$ is typically learned jointly with Z and therefore $\bar{\delta}(\hat{\mathbb{P}})$ may itself depend on $I(X;Z)$. Thus we cannot yet say that this relationship is truly linear, and we certainly cannot yet say that it is tight. Before we can make those claims, we will need to study $\bar{\delta}(\hat{\mathbb{P}})$ explicitly. We will begin with a 'sanity-check' lemma. This lemma shows us that $\bar{\delta}(\hat{\mathbb{P}})$ does at least converge with the convergence of a typical neural classifier loss function. It arises from an application of Pinsker's inequality [21].

Lemma 5. *Suppose that $H(Y|X) = 0$. Then:*

$$\bar{\delta}(\hat{\mathbb{P}}) \leq \sqrt{\frac{1}{2}H_{\mathbb{P},\hat{\mathbb{P}}}(Y|X)} \tag{3.32}$$

where $H_{\mathbb{P},\hat{\mathbb{P}}}(Y|X)$ is the conditional cross entropy between $\mathbb{P}$ and $\hat{\mathbb{P}}$, i.e. the usual cross entropy loss function.

Proof.

$$\begin{aligned}
\bar{\delta}(\hat{\mathbb{P}}) &= \int \delta_{TV}(\mathbb{P}_{Y|X}, \hat{\mathbb{P}}_{Y|X}) d\mathbb{P}_X \\
&\leq \int \sqrt{\frac{1}{2}\mathcal{D}_{KL}\left[\mathbb{P}_{Y|X} \,||\, \hat{\mathbb{P}}_{Y|X}\right]} d\mathbb{P}_X \\
&\leq \sqrt{\int \frac{1}{2}\mathcal{D}_{KL}\left[\mathbb{P}_{Y|X} \,||\, \hat{\mathbb{P}}_{Y|X}\right] d\mathbb{P}_X} \\
&= \sqrt{\frac{1}{2} H_{\mathbb{P},\hat{\mathbb{P}}}(Y|X)}
\end{aligned} \tag{3.33}$$

□

This lemma is particularly applicable when we are estimating our cross entropy error on a validation set, as we can then take $\mathbb{P}$ in this lemma to be the empirical measure corresponding to the validation or training sample, in which we are almost certain to have $H(Y|X) = 0$. In this sense Lemma 5 can bound such empirical estimates of $\bar{\delta}(\hat{\mathbb{P}})$.

3.2 Finite Bounds for Discrete Spaces

We will next study $\bar{\delta}(\hat{\mathbb{P}})$ in the case of a discrete feature space before moving on to a more general case. In this case, we have a non-asymptotic upper bound that is $O\left(\sqrt{\frac{|\mathcal{X}||\mathcal{Y}|}{2m}}\right)$.

Theorem 2. *Let $0 < \nu < 1$. Let X be discrete. Suppose further that we can choose which data points to label. If we then choose to label $m(x_i) = \lceil mp(x) \rceil$ points for each x, then:*

$$\bar{\delta}(\hat{\mathbb{P}}) \leq \sqrt{\frac{|\mathcal{Y}||\mathcal{X}| log\left(\frac{m}{|\mathcal{X}|} + 2\right) + log(\frac{1}{\nu})}{2m}} \tag{3.34}$$

holds with probability at least $1 - \nu$

Proof. Let

$$\mathcal{P} = \left\{ \begin{bmatrix} \frac{k_1^1}{m(x_1)} & \frac{k_2^1}{m(x_1)} & \cdots & \frac{k_{|\mathcal{Y}|}^1}{m(x_1)} \\ \frac{k_1^2}{m(x_2)} & \frac{k_2^2}{m(x_2)} & \cdots & \frac{k_{|\mathcal{Y}|}^2}{m(x_2)} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{k_1^{|\mathcal{X}|}}{m(x_{|\mathcal{X}|})} & \frac{k_2^{|\mathcal{X}|}}{m(x_{|\mathcal{X}|})} & \cdots & \frac{k_{|\mathcal{Y}|}^{|\mathcal{X}|}}{m(x_{|\mathcal{X}|})} \end{bmatrix} : k_j^i \geq 0 \mathbb{Z}, \sum_{j=1}^{|\mathcal{Y}|} k_j^i = m(x_i) \; \forall i = 1, 2, \cdots, |\mathcal{X}| \right\} \tag{3.35}$$

That is, if $Q \in \mathcal{P}$, then the i^{th} row of Q is an estimated conditional probability distribution that is obtainable by the sampling procedure defined in the statement of the theorem. Here k_j^i is to be interpreted as the number of occurrences of class j when sampling from $p(\cdot|i)$. We call an element of $\mathcal{P}$ a multi-type. Multi types are slightly different from the well known conditional types. To each multi-type Q, there is an associated distribution $q(y|x)$ given by $q(y = j|x = i) = Q_{ij}$. We will abuse notation by using these somewhat interchangeably when the corresponding multi-type is clear. We will further often write terms like $p(j|i)$ as a stand in for $p(y = j|x = i)$.

The next few steps develop a Large Deviations inequality for the multi-type object. It follows closely to the standard development of Large Deviations theory for discrete random variables [24] (Chapter 2).

First, we wish to find the probability that our sampling procedure will yield the multi-type Q. To this end, let

$$T(Q) = \Big\{ (y_1^1, y_2^1, \cdots, y_{m(x_1)}^1), \cdots, (y_1^{|\mathcal{X}|}, y_2^{|\mathcal{X}|}, \cdots, y_{m(x_{|\mathcal{X}|})}^{|\mathcal{X}|}) : \frac{1}{m(x_i)} \sum_{l=1}^{m(x_i)} \delta(y_l^i = j) = Q_{ij} \Big\} \tag{3.36}$$

That is, $T(Q)$ is the set of sequences which yield the multi-type Q. Let s be a random

sequence of samples from the true distribution $p(y|x)$. Let $\hat{p}$ be the corresponding multi-type from s. Then:

$$\begin{aligned}\mathbb{P}(\hat{p} = Q) = \mathbb{P}(s \in T(Q)) &= \sum_{s \in T(Q)} \prod_{i=1}^{|\mathcal{X}|} \prod_{j=1}^{|\mathcal{Y}|} p(j|i)^{m(x_i)Q_{ij}} \\ &= |T(Q)| e^{\sum_i \sum_j m(x_i) Q_{ij} log\ p(j|i)} = |T(Q)| e^{\sum_i \lceil mp(x_i) \rceil \sum_j q(j|i) log\ p(j|i)} \\ &= |T(Q)| e^{-\sum_i \lceil mp(x_i) \rceil [\mathcal{D}_{KL}(q(\cdot|i)\ ||\ p(\cdot|i)) + H(q(\cdot|i))]} \end{aligned} \tag{3.37}$$

We next wish to bound $|T(Q)|$. To do so, let s_q be a random sequence of samples from the distribution $q(y|x)$. Let $\hat{q}$ be the multi-type obtained from s_q. By then applying Equation (3.37) to $\hat{q}$, we have:

$$1 \geq \mathbb{P}(\hat{q} = Q) = |T(Q)| e^{-\sum_i \lceil mp(x_i) \rceil H(q(\cdot|i))} \tag{3.38}$$

so $|T(Q)| \leq e^{\sum_i \lceil mp(x_i) \rceil H(q(\cdot|i))}$. Thus, in total, we have:

$$\begin{aligned}\mathbb{P}(\hat{p} = Q) &\leq e^{-\sum_i \lceil mp(x_i) \rceil [\mathcal{D}_{KL}(q(\cdot|i)\ ||\ p(\cdot|i))]} \leq e^{-m \sum_i p(x_i) [\mathcal{D}_{KL}(q(\cdot|i)\ ||\ p(\cdot|i))]} \\ &= e^{-m \mathbb{E}_{p(x)} [\mathcal{D}_{KL}(q(\cdot|x)\ ||\ p(\cdot|x))]} \end{aligned} \tag{3.39}$$

Now, let $\epsilon > 0$ and let $\Gamma = \left\{ Q \in \mathcal{P} : \mathbb{E}_{p(x)} \left[\frac{1}{2} \sum_{j=1}^{|\mathcal{Y}|} |p(j|x) - q(j|x)| \right] \geq \epsilon \right\}$ Then:

$$\mathbb{P}(\hat{p} \in \Gamma) = \sum_{Q \in \Gamma \cap \mathcal{P}} \mathbb{P}(\hat{p} = Q) \leq |\mathcal{P}| e^{-m \inf_{\gamma \in \Gamma} \mathbb{E}_{p(x)} [\mathcal{D}_{KL}(\gamma(\cdot|x)\ ||\ p(\cdot|x))]} \tag{3.40}$$

But:

$$|\mathcal{P}| \leq \prod_{i=1}^{\mathcal{X}} (\lceil mp(x_i) \rceil + 1)^{|\mathcal{Y}|} \leq \prod_{i=1}^{\mathcal{X}} (mp(x_i) + 2)^{|\mathcal{Y}|}$$
$$\leq \left(\frac{\sum_x mp(x) + 2}{|\mathcal{X}|} \right)^{|\mathcal{X}||\mathcal{Y}|} = \left(\frac{m}{|\mathcal{X}|} + 2 \right)^{|\mathcal{X}||\mathcal{Y}|} \tag{3.41}$$

Further, by Pinsker's inequality [21], we have for all $\gamma \in \Gamma \cap \mathcal{P}$ that:

$$\mathbb{E}_{p(x)} \left[\mathcal{D}_{KL} \left(\gamma(\cdot|x) \,||\, p(\cdot|x) \right) \right] \geq 2\epsilon^2 \tag{3.42}$$

Thus $\inf_{\gamma \in \Gamma} \mathbb{E}_{p(x)} \left[\mathcal{D}_{KL} \left(\gamma(\cdot|x) \,||\, p(\cdot|x) \right) \right] \geq 2\epsilon^2$. We then have:

$$\mathbb{P}(\hat{p} \in \Gamma) \leq \left(\frac{m}{|\mathcal{X}|} + 2 \right)^{|\mathcal{X}||\mathcal{Y}|} e^{-2m\epsilon^2} \tag{3.43}$$

Finally, we desire $\mathbb{P}(\hat{p} \in \Gamma) \leq \nu$. This will be guaranteed as long as we have that $\nu \geq \left(\frac{m}{|\mathcal{X}|} + 2 \right)^{|\mathcal{X}||\mathcal{Y}|} e^{-2m\epsilon^2}$, and this occurs so long as:

$$\epsilon \geq \sqrt{\frac{|\mathcal{Y}||\mathcal{X}| log(\frac{m}{|\mathcal{X}|} + 2) + log(\frac{|\mathcal{X}|}{\nu})}{2m}} \tag{3.44}$$

completing the proof. □

3.2.1 Bounding $\bar{\delta}(\hat{\mathbb{P}})$ - Setting

Finally, we will derive a rate of decrease for $\bar{\delta}(\hat{\mathbb{P}})$ in a general continuous learning algorithm. Our setup will involve defining a learning algorithm as a continuous map from a special topology on input probability measures on $\mathcal{X} \times \mathcal{Y}$ to conditional probability functions. This is basically to say that, given a training dataset (i.e. an empirical measure on $\mathcal{X} \times \mathcal{Y}$), we have a well-behaved way of obtaining the corresponding $\hat{p}_\nu(y|x)$. This is just slightly generalized

so that we can consider any input measure (empirical or not) as a 'training dataset'. We begin by reviewing that special topology, and then we will construct the topology that we will place on our output conditional probability distributions.

Definition 3. *Let M_1 denote the set of Borel probability measures on $\mathcal{X} \times \mathcal{Y}$. Then the τ-topology [24] (page 263) is the topology generated by the family of sets given by $W_{f,r,c} = \{\nu : |\int f d\nu - r| < c\}$ where the index f runs over all bounded Borel measurable functions on $\mathcal{X} \times \mathcal{Y}$ to the reals, the index r runs over the reals, and the index c runs over the positive reals. If we restrict the indexing set of f to the set of bounded* continuous *functions, then we get the weak topology $\mathcal{W}$, which is strictly coarser than the τ-topology.*

Definition 4. *Let $\Sigma_{|\mathcal{Y}|}$ be the probability simplex in $|\mathcal{Y}|$ dimensions. Let $L^1(\mathcal{X})$ denote the space of absolutely integrable functions from $\mathcal{X}$ to $\mathbb{R}$ with norm $\|f\|_{L_1} = \int |f| d\mathbb{P}_{\mathbb{X}}$. Let $L^1(\mathcal{X})^{|\mathcal{Y}|}$ denote the product space on $L^1(\mathcal{X})$, consisting of functions from $\mathcal{X}$ to $\mathbb{R}^{|\mathcal{Y}|}$ which are absolutely integrable in each output dimension, and with norm*

$$\|f\|_{L_1^{|\mathcal{Y}|}} = \frac{1}{2} \int \sum_y |f(x,y)| d\mathbb{P}_{\mathbb{X}} \tag{3.45}$$

Finally, let $L^1(\mathcal{X}, \Sigma_{|\mathcal{Y}|})$ denote the subspace of $L^1(\mathcal{X})^{|\mathcal{Y}|}$ given by the set of functions whose co-domain is $\Sigma_{|\mathcal{Y}|}$.

The topology we've placed on $L^1(\mathcal{X}, \Sigma_{|\mathcal{Y}|})$ is metrized by the conditional total variation function that we've been working with. With these topologies defined, we will restrict ourselves to the study of algorithms which act as continuous maps between these topologies. This essentially requires that, when our training datasets are very similar (e.g. moving one training point to a point within a distance ϵ from the original), our algorithm will return very similar output functions in terms of conditional total variation. Thus this condition is somewhat related to algorithmic stability [43], though not completely equivalent.

We will obtain two bounds on $\bar{\delta}(\hat{\mathbb{P}})$ in the remains of this chapter. The first is asymptotic, and applies when we have continuity from the τ-topology. The second is non asymptotic, and applies when we further have continuity from the weak topology. We will next show that gradient descent algorithms, under mild conditions, achieve these continuities.

Theorem 3. *Let Θ denote a normed parameter space and let $\mathcal{L} : \mathcal{X} \times \mathcal{Y} \times \Theta \to \mathbb{R}$ denote a loss function which is integrable in $\mathcal{X} \times \mathcal{Y}$ for each $\theta \in \Theta$, which is differentiable with respect to θ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and whose θ-gradients yield bounded continuous functions on $\mathcal{X} \times \mathcal{Y}$ when evaluated at each point $\theta \in \Theta$. Suppose further that our parameter space admits lipschitz-continuous outputs for each (x, y). That is, $|p_{\theta_1}(y|x) - p_{\theta_2}(y|x)| < L\|\theta_1 - \theta_2\|$ $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$. Then gradient descent applied to the empirical risk minimization of $\mathcal{L}$, with a fixed initiation $\theta^{(0)}$ and which proceeds for a fixed number of iterations, is continuous from $(M_1, \mathcal{W})$ to $L^1(\mathcal{X}, \Sigma_{|\mathcal{Y}|})$.*

If we relax the condition that the θ gradients of $\mathcal{L}$ be bounded continuous functions on $\mathcal{X} \times \mathcal{Y}$ when evaluated at each point $\theta \in \Theta$ to just bounded measurable functions, then this algorithm is still continuous from (M_1, τ) to $L^1(\mathcal{X}, \Sigma_{|\mathcal{Y}|})$.

Proof. The assumptions on $\mathcal{L}$ allow us to differentiate (with respect to θ) under the integral sign. Let α_k denote the step size of the k^{th} iteration. Let $\nu^* \in M_1$. We proceed by induction on the number of iterations.

Let $\epsilon > 0$. Let $\delta_1 = \frac{2\epsilon}{L\alpha_1|\mathcal{Y}|}$. Let $\nu^* \in M_1$ and let ν be contained in the open set of the weak topology given by $\{\nu : |\int \nabla_{\theta^{(0)}} d\nu - \int \nabla_{\theta^{(0)}} d\nu^*| < \delta_1\}$ (which clearly contains ν^*). Let $\theta_*^{(1)}$ denote the parameter chosen after one gradient update when training on ν^*, and let $\theta^{(1)}$ denote the parameter chosen after one gradient update when training on ν. Then:

$$\|\theta_*^{(1)} - \theta^{(1)}\| = \left\| \alpha_1 \left(\int \nabla_{\theta^{(0)}} d\nu - \int \nabla_{\theta^{(0)}} d\nu^* \right) \right\| \leq \alpha_1 \delta_1 \tag{3.46}$$

so

$$\frac{1}{2}\int \sum_{y} \|p_{\theta_*^{(1)}}(y|x) - p_{\theta^{(1)}}(y|x)\| d\mathbb{P}_X \leq \frac{L|\mathcal{Y}|\alpha_1\delta_1}{2} = \epsilon \tag{3.47}$$

and so the hypothesis is true if our algorithm consists of one iteration.

Suppose that the hypothesis is true when we use $(k-1)$ iterations. Let $\epsilon > 0$. Let $\delta_{k-1} = \frac{\epsilon}{L|\mathcal{Y}|}$ and let $\delta_k = \frac{\epsilon}{L|\mathcal{Y}|\alpha_k}$. Chose an open set U of the weak topology such that $\|\theta_*^{(k-1)} - \theta_c^{(k-1)}\| \leq \delta_k$ when $\nu_c \in U$ which is possible by the induction hypothesis, and where $\theta_*^{(k-1)}$ and $\theta_c^{(k-1)}$ denote the chosen parameters after iteration $k-1$ of the gradient descent when trained on ν^* and ν_c. Let $\nu \in U \cap \{\nu : |\int \nabla_{\theta^{(k-1)}} d\nu - \int \nabla_{\theta^{(k-1)}} d\nu^*| < \delta_k\}$. Then by the triangle inequality:

$$\|\theta_*^{(k)} - \theta^{(k)}\| \leq \delta_{k-1} + \alpha_k\delta_k \tag{3.48}$$

so the conditional total variation between $p_{\theta_*^{(k)}}(y|x)$ and $p_{\theta^{(k)}}(y|x)$ is less than or equal to $\frac{L|\mathcal{Y}|(\delta_{k-1}+\alpha_k\delta_k)}{2}$ which is equal to ϵ.

For the final statement, note that all of the above open sets in the $\mathcal{W}$-topology used in this proof remain open sets in the τ-topology when we relax the conditions of $\mathcal{L}$. This completes the proof. □

3.2.2 Bounding $\bar{\delta}(\hat{\mathbb{P}})$ - The Asymptotic Case

We now wish to bound the conditional total variation of an estimated model against the true model when we use such a general learning algorithm in our setting. We will re-label $\bar{\delta}(\hat{\mathbb{P}})$ to $\bar{\delta}(\mathbb{P}_f)$ to emphasize that our estimated model is coming from such an algorithm. We then have the following asymptotic theorem on the rate of decay for $\bar{\delta}(\mathbb{P}_f)$. This will apply whenever we have continuity from the τ-topology in our algorithm, and will be used in our

non-asymptotic specialization that follows. We will use two final lemmas in both of those proofs.

Lemma 6. *Let $(\Omega, \mathcal{F}, \mu)$ be a probability space and let $h : \Omega \to \mathbb{R}$ be bounded and measurable. Let $\mathcal{G}$ denote the set of non-negative measurable functions with expectation 1. Then $\inf_{g \in \mathcal{G}} \mathbb{E}\left[g \cdot (h + log\ g)\right] = -log\ \mathbb{E}\left[e^{-h(\omega)}\right]$.*

Proof. This infinum can be found by the following Lagrangian:

$$\mathcal{L} = \mathbb{E}\left[g \cdot (h + log\ g)\right] + \lambda\left(\mathbb{E}\left[g\right] - 1\right) \tag{3.49}$$

(we will see that we don't need to worry about the $g(\omega) \geq 0$ constraints because the solution to the lagrangian we just wrote will yield a function g in which those constraints are not tight). The functional derivative of this Lagrangian is $h(\omega) + log\ g(\omega) + 1 + \lambda$. Fixing this to zero yields $g(\omega) = e^{-\lambda} e^{-(h(\omega)+1)}$. Setting λ through normalization then yields $g(\omega) = \frac{1}{W} e^{-(h(\omega)+1)}$ where $W = \mathbb{E}\left[e^{-(h(\omega)+1)}\right]$. Plugging this solution into our objective yields $-1 - log\ W = -log\ \mathbb{E}\left[e^{-(h(\omega)+1)}\right] - 1$. Since our objective function was a strictly convex functional with a positive second variation given by $\frac{1}{g(\omega)}$, this is a minimizer. □

Lemma 7. *Let $(\Omega, \mathcal{F}, \mu)$ be a probability space and let $f : \Omega \to \mathbb{R}$ be bounded and measurable with $Range(f) \subseteq [0, 1]$. Then $log\ \left(\mathbb{E}\left[e^{-2f^2}\right]\right) \leq -2\mathbb{E}\left[f\right]^2$.*

Proof. This follows from reference [61] (Theorem 1) with $\phi = -log(\cdot)$ while replacing $h(x; \mu)$ with $\phi''(x)/2 = \frac{1}{2x^2}$. Denote $Y = e^{-2f^2}$. The range of Y is a subset of $[e^{-2}, 1]$. On this set, the supremum of $\phi''(x)/2$ is $\frac{1}{2}$. Thus $log\left(\mathbb{E}\left[Y\right]\right) \leq \mathbb{E}\left[log(Y)\right] + \frac{1}{2} Var\left[Y\right]$. But $Var\left[e^{-2f^2}\right] \leq 4Var[f^2] \leq 4Var[f]$ (because f has range bounded by $[0, 1]$). We thus have $log\left(\mathbb{E}\left[e^{-2f^2}\right]\right) \leq -2\mathbb{E}\left[f^2\right] + 2Var\left[f\right]$. This completes the proof since $Var\left[f\right] = \mathbb{E}[f^2] - \mathbb{E}[f]^2$. □

Theorem 4. *Let $\epsilon \in (0,1)$, and let $0 < \zeta < 1$. If $\mathcal{F}$ is a continuous learning algorithm from (M_1, τ) to $L^1(\mathcal{X}, \Sigma_{|\mathcal{Y}|})$ such that, for any $\nu \in M_1$, the total variation between $\mathcal{F}\nu$ and $\nu_{y|x}$ is smaller than the total variation between $\mathcal{F}\nu$ and $p_{y|x}$ at any point in the support of ν. Suppose further that the 'training' total variation, $\mathbb{E}_\nu \left[\frac{1}{2}\sum_y |\nu_{y|x} - \mathcal{F}\nu|\right]$, is bounded above by ζ. Then:*

$$\underset{m\to\infty}{\textit{limsup}}\ \frac{1}{m} log\ \mathbb{P}^m(\bar{\delta}(\mathbb{P}_f) \geq \epsilon) \leq 4\zeta - 2\epsilon^2 \tag{3.50}$$

where $\mathbb{P}^m$ is the probability measure on M_1 induced by the sampling of m data-points on $\mathcal{X} \times \mathcal{Y}$.

Proof. For notational convenience, we will denote as $\delta_\nu(x)$ the conditional total variation between $p(y|x)$ and $(\mathcal{F}\nu)(y|x)$ for a fixed x.

We will first need to show that the map $\bar{\delta} : M_1 \to \mathbb{R}$, given by $\nu \mapsto \mathbb{E}_{\mathbb{P}_X}[\delta_\nu]$ is continuous from the τ-topology to the Euclidean topology. This is trivial since $\mathbb{E}_{\mathbb{P}_X}[\delta_\nu]$ is just the composition of $\mathcal{F}$, which was assumed continuous, with the fixed-point distance function $d(\cdot, p_{y|x}(y|x))$ defined over $L^1(\mathcal{X}, \Sigma_{|\mathcal{Y}|})$.

Now, let $\Gamma = \{\nu \in M_1 : \mathbb{E}_{\mathbb{P}_X}[\delta_\nu] \geq \epsilon\}$. By the above continuity and by the fact that $[\epsilon, 1]$ is closed in $\mathbb{R}$, we have that Γ is closed. Then, by Sanov's Theorem [24]:

$$\underset{m\to\infty}{\text{limsup}}\ \frac{1}{m} log\ \mathbb{P}^m(\mathbb{P}_f \in \Gamma) \leq -\inf_{\nu\in\Gamma} \mathcal{D}_{KL}(\nu||p(x,y)) \tag{3.51}$$

We thus wish to lower bound $\mathcal{D}_{KL}(\nu||p(x,y))$ over Γ. We begin by decomposing $\frac{d\nu}{d\mathbb{P}}$ into $\frac{d\nu_x}{d\mathbb{P}_x}\frac{\nu_{y|x}}{p_{y|x}}$. Where ν_x and $\mathbb{P}_x$ are the marginal distributions of ν and $p(x,y)$ on $\mathcal{X}$. We are guaranteed that the functions and $\nu_{y|x}$ exist on the support of ν_x since y is discrete. The KL-divergence then becomes: $\mathcal{D}_{KL}(\nu||p(x,y)) = \mathbb{E}_{\mathbb{P}_X}\left[\frac{d\nu_x}{d\mathbb{P}_x}(\tilde{h} + log\frac{d\nu_x}{d\mathbb{P}_x})\right]$ where $\tilde{h} \triangleq \sum_y \nu_{y|x} log\ \frac{\nu_{y|x}}{p_{y|x}}$ is bounded below (via Pinsker's inequality) by $\left(\sum_y |p_{y|x} - \nu_{y|x}|\right)^2$,

which itself is bounded below by $2\left(\sum_y |p_{y|x} - \mathcal{F}\nu| - \sum_y |\nu_{y|x} - \mathcal{F}\nu|\right)^2$ because the absolute value of the second term in this expression is smaller than that of the first term for each point in the support of ν. The first term is just the function δ_ν defined at the start of this proof. We will call the second term δ_ν^t. We can lower bound this expression one more time with $2\delta_\nu^2 - 4\delta_\nu^t$. We are left with:

$$\mathcal{D}_{KL}(\nu||p(x,y)) \geq \mathbb{E}_{\mathbb{P}_X}\left[\frac{d\nu_x}{d\mathbb{P}_x}(2\delta_\nu^2 + log\frac{d\nu_x}{d\mathbb{P}_x})\right] - 4\mathbb{E}_\nu\left[\delta_\nu^t\right] \tag{3.52}$$

We will bound these two remaining terms separately. The second is taken care of in this theorem's hypothesis, being bounded below by -4ζ. For the latter, we can combine Lemmas 6 and 7 to obtain a lower bound of $2\epsilon^2$ (since $\nu \in \Gamma$). □

3.2.3 Bounding $\bar{\delta}(\hat{\mathbb{P}})$ - The Non-Asymptotic Case

The previous theorem gives us:

$$\mathbb{P}^m(\bar{\delta}(\mathbb{P}_f) \geq \epsilon) \leq e^{m(4\zeta - 2\epsilon^2) + o(m)} \tag{3.53}$$

where $o(m)$ refers to any terms such that $\lim_{m\to\infty} \frac{o(m)}{m} = 0$. We will need to study $o(m)$ since it's somewhat of an unknown here, and may be large for small m. The next theorem, which is non-asymptotic, will take care of this when $\mathcal{F}$ is continuous from the weak topology.

Theorem 5. *Take all assumptions from Theorem 4, but remove the assumption that $\mathcal{F}$ be a continuous map from (M_1, τ) to $L^1(\mathcal{X}, \Sigma_{|\mathcal{Y}|})$ and assume it is instead continuous from $(M_1, \mathcal{W})$. Suppose further that $\mathcal{X}$ is compact, and that $p(x)$ has full support with density $p(x,y) > 0$ everywhere. Then there exists a function $k(m') : \mathbb{Z}^+ \to \mathbb{R}$ with $k(m') \leq \sqrt{m'}$*

such that:

$$\mathbb{P}^m(\bar{\delta}(\mathbb{P}_f) \geq \epsilon) \leq \inf_{m' \in \mathbb{Z}^+} 2^{m'|\mathcal{Y}|} e^{-2m\left(\left(\epsilon - 2\frac{k(m')}{\sqrt{m'}}\right)^2 - 4\zeta\right) + 2\frac{k(m')}{\sqrt{m'}}} \tag{3.54}$$

(A more detailed description of $k(m')$, from which we can discover more of its properties, is contained in the proof).

Proof. Let the notations δ_ν and Γ be defined as they were in the proof of Theorem 4.

Let $E(S_{m'}, k(m'))$ constitute a family of conditions, indexed first by samples of m' points of $\mathcal{X}$ and second by functions $\mathbb{Z}^+ \to \mathbb{R}$, which constitute that:

$$|\mathbb{E}_{p(x)}[\delta_\nu] - \mathbb{E}_{S_{m'}}[\delta_\nu]| \leq \frac{k(m')}{\sqrt{m'}} \tag{3.55}$$

where the second expectation is the monte-carlo estimate over the indexed sample.

Let the sets $\Gamma(S_{m'}, i)$, indexed first over samples of $\mathcal{X}$ consisting of m' points and second over the set $1, 2, \cdots, 2^{m'|\mathcal{Y}|}$, be given by:

$$\Gamma(S_{m'}, i) = \{h : \mathbb{E}_{p(x)}[\delta_h] \geq \epsilon,\ \mathcal{F}h(y|x_j) \geq / \leq p_{y|x}(y|x_j)\} \tag{3.56}$$

(where the $x'_j s$ run over the sampled points in $S_{m'}$ and i runs over the possible choices of $\geq / \leq$). Let denote the family of conditions:

$$F(S'_m, i, k(m')) = \{\nu : \mathbb{E}_{S_{m'}}[\delta_\nu] \geq \epsilon - \frac{k(m')}{\sqrt{m'}},\ \mathcal{F}\nu(y|x_j) \geq / \leq p_{y|x}(y|x_j)\} \tag{3.57}$$

where the x_j run over the sampled points and the choices of $\geq$ and $\leq$ correspond to those of Γ^i. Let $G(S_{m'}, i)$ denote the condition on measures $\mu \in M_1$ such that there exists a measure $\mu' \in \Gamma(S_{m'}, i)$ with $\mu'_{y|x} = \mu_{y|x}$. Note that $E(S_{m'}, k(m')) \cap G(S_{m'}, i) \subseteq F(S'_m, i, k(m'))$.

Let M denote the vector space of finite signed measures on $\mathcal{X} \times \mathcal{Y}$ endowed with the weak topology. For any probability measure $\nu'_x \in M_1(\mathcal{X})$, let $R^{\nu'_x}$ be the subspace of measures with marginal distribution ν'_x. Let $R_1^{\nu'_x}$ be the subset of $R^{\nu'_x}$ consisting of probability measures. Define a linear map on $R_1^{\nu'_x}$, denoted $\mathcal{C}_{\nu'_x}$, which takes ν' to its disintegration $\nu'_{y|x}$.

Let $f_{\nu'_x} : M_1 \times \mathcal{C}_{\nu'_x} R_1^{\nu'_x}$ denote the family of real valued function (indexed by $M_1(\mathcal{X})$) taking $(\nu, \nu'_{y|x})$ to the value $\mathbb{E}_{\nu_x}\left[\sum_y \nu_{y|x} log \frac{\nu'_{y|x}}{p_{y|x}} + log\ \frac{d\nu_x}{d\mathbb{P}_X}\right]$, which is to be taken as infinite when the support of ν'_x is not a superset of the support of ν_x, and is further infinite when ν_x is not absolutely continuous with respect to $p(x)$. Note that each $f_{\nu'_x}(\cdot, a)$ is convex and continuous in the weak topology for each fixed a (as $p(x) > 0$ and $p_{y|x} > 0$ everywhere by the theorem's hypothesis), and each $f_{\nu'_x}(b, \cdot)$ is concave and continuous for each fixed b.

Now, since $\mathcal{X} \times \mathcal{Y}$ is compact, M_1 is compact in the weak topology. Then for any ν'_x, $R_1^{\nu'_x}$ is compact (being a closed subset of a compact space). Then $\mathcal{C}_{\nu'_x} R_1^{\nu'_x}$ is compact and convex. We also have that the subsets $G(S_{m'}, i)$, $E(S_{m'}, k(m'))$, and $F(S_{m'}, i, k(m'))$ are all closed, and therefore compact. We also have convexity in $F(S_{m'}, i, k(m'))$, but not in the other two.

Arbitrarily pick some $\nu''_x \in M_1$ with full support and denote f as $f_{\nu''_x}$ as f. Let $r(S_{m'}, i, k(m'))$ denote the minimum of the expression $f(a, a_{y|x})$ over the intersection of sets given by $K(S_{m'}, i) \cap E(k(m')) \cap F(S_{m'}, i, k(m'))$ and denote the minimizer as $a(S_{m'}, i, k(m'))$. The image of the map $f(\cdot, a(S_{m'}, i, k(m')))$ is a compact subset of $\mathbb{R}$ - i.e. a closed and bounded interval $\mathcal{I}(S_{m'}, i, k(m'))$. Let $\tilde{\mathcal{I}}(S_{m'}, k(m'))$ denote the union of these intervals over the finite indices i. Cover this interval with a family of subintervals $\tilde{\mathcal{I}}(S_{m'}, k(m'), j)$ of size $\frac{k(m')}{\sqrt{m'}}$.

We will now fix $k(m')$ to be the smallest number such that *there exists* a sample $S^*_{m'}$ in which both $G(S^*_{m'}, i) \cap E(S^*_{m'}, k(m')) \neq \emptyset$ for all i in which $G(S^*_{m'}, i) \neq \emptyset$ and $\mathcal{I}(S^*_{m'}, k(m'), j) \cap E(S^*_{m'}, k(m')) \neq \emptyset$ for all j in which $\tilde{\mathcal{I}}(S^*_{m'}, k(m'), j) \neq \emptyset$. Such a

$k(m')$ exists, and is less than or equal to $\sqrt{m'}$ since $E(S_{m'}, \sqrt{m'})$ is all of M_1. Fix $S_{m'}$ to any of the samples that we just established the existence of. We will drop the notations $S_{m'}$ and $k(m')$ from the notation for any conditions referring to them from now on.

Now, denote as $C_b(\mathcal{X})$ the set of bounded continuous functions from $\mathcal{X}$ to $\mathbb{R}$ and construct a family of maps $\mathcal{G}_{\lambda,\nu'} : M_1 \to \mathbb{R}$ indexed over $\lambda \in C_b(\mathcal{X})$ and $\nu' \in M_1$ which takes $\nu \in M_1$ to $\mathbb{E}_\nu \left[m log \frac{\nu'_{y|x}}{p_{y|x}} + m\lambda \right]$. Then for any empirical $L_m \in \Gamma(i)$ corresponding to a sample of m points, we have that $\mathcal{G}_{\lambda,\nu'} L_m \geq \inf_{\nu \in \Gamma(i)} \mathcal{G}_{\lambda,\nu'} \nu$ for all λ, ν'. Thus the probability that L_m is in $\Gamma(i)$ is bounded above by the probability that $\mathcal{G}_{\lambda,\nu'} L_m - \inf_{\nu \in \Gamma(i)} \mathcal{G}_{\lambda,\nu'} \nu \geq 0$. Then by Chernoff's inequality, we have that $\frac{1}{m} log\, \mathbb{P}^m \left(L_m \in \Gamma(i) \right)$ is bounded above by:

$$\frac{1}{m} log\, \mathbb{E} \left[e^{m \mathbb{E}_{L_m} \left[log \frac{\nu'_{y|x}}{p_{y|x}} + \lambda \right]} \right] - \inf_{\nu \in \Gamma(i)} \mathbb{E}_\nu \left[log \frac{\nu'_{y|x}}{p_{y|x}} + \lambda \right] \tag{3.58}$$

where the first expectation is taken over $\mathbb{P}^m$.

The first term can be reduced to $log\, \mathbb{E}_{p(x)} \left[e^\lambda \right]$. Optimizing over λ yields a bound of

$$-\sup_\lambda \inf_{\nu \in \Gamma(i)} \mathbb{E}_\nu \left[log \frac{\nu'_{y|x}}{p_{y|x}} \right] + \mathbb{E}_\nu \left[\lambda \right] - log(\mathbb{E}_{p(x)} \left[e^\lambda \right]) \tag{3.59}$$

We will denote as $\Gamma^i_{y|x}$ the set of conditional probability functions $\nu_{y|x}$ such that there exists $\nu \in \Gamma(i)$ with disintegration given by $\nu_{y|x}$. We will also introduce a function, denoted as, $g_{\nu'}(\nu_{y|x}, \mu_x)$ defined on $\Gamma^i_{y|x} \times M_1(\mathcal{X})$ which yields $\mathbb{E}_{\mu_x \nu_{y|x}} \left[log \frac{\nu'_{y|x}}{p_{y|x}} \right]$ when the support of the latter argument is equal to the domain of the former, and is infinite otherwise. Note that g is convex and lower-semicontinuous in μ_x for fixed $\nu_{y|x}$ since it is linear in the convex subset $\{\mu_x \in M_1(\mathcal{X}) : supp(\mu_x) = Dom(\nu_{y|x})\}$ and infinite outside of this subset. Finally, we will define the function $h : M_1(\mathcal{X}) \times C_b(\mathcal{X}) \to \mathbb{R}$ which takes (μ_X, λ) to $\mathbb{E}_{\mu_x} \left[\lambda \right] - log(\mathbb{E}_{p(x)} \left[e^\lambda \right])$. This function is concave in λ, convex in μ_x, and lower semicon-

tinuous in μ_x [24]. Then (3.59) is upper bounded by:

$$-\sup_{\lambda \in C_b \nu_{y|x} \in \Gamma^i_{y|x}} \inf_{\mu_x \in M_1(\mathcal{X})} g_{\nu'}(\nu_{y|x}, \mu_x) + h(\mu_x, \lambda) \tag{3.60}$$

Note also that the objective function of this expression is decoupled for $\nu_{y|x}$ and λ. We can thus swap the supremum with the first infinum. But then inside the first infinum, we are left with an objective function in which a minimax theorem applies [94] because $M_1(\mathcal{X})$ is compact and convex in the weak topology when $\mathcal{X}$ is compact, and so we can swap the supremum with the second infinum as well. Since the first term does not depend on λ, we can then consider for each fixed μ_x the expression $\sup_\lambda \; h(\mu_x, \lambda)$. But the supremum of this function over $\lambda \in C_b(\mathcal{X})$ is none other than the KL divergence between μ_x and $p(x)$ [27]. We are thus left with a full upper bound of (now optimizing over $\nu'_{y|x} \in \mathcal{C}_{\nu''_x} R_1^{\nu''_x}$):

$$-\sup_{\nu'} \underset{\nu \in G(i)}{f} (\nu_{y|x}, \nu'_{y|x}) \tag{3.61}$$

We would be able to swap the supremum and infinum if our feasible set were convex and compact. This is true for our search space over ν', but not for $G(i)$. Our goal is to then transform $G(i)$ into $F(i)$, which is convex, with corresponding error terms included. This can be done by tightening $G(i)$ to $G(i) \cap E$ and then relaxing that set to $F(i)$, this will incur some error, but if we end up choosing $\nu'_{y|x}$ to be the disintegration of $a(i)$, then this error will be bounded by $\frac{k(m')}{\sqrt{m'}}$.

With our feasible set now being $F(i)$, we can swap the supremum and infinum, and then pick $\nu'_{y|x}$ to be equal to $\nu_{y|x}$ on the support of ν, and arbitrary elsewhere. The objective function is then just the minimum KL divergence over $F(i)$, which we know how to deal with due to the proof of Theorem 4. Minimizing then gives us $\nu_{y|x} = \nu'_{y|x}$ both given by the disintegration of $a(i)$, and with the objective function bounded by $\inf_{\nu \in F(i)} 2\mathbb{E}_{p(x)} \left[\delta_\nu\right]^2 - 4\zeta$.

If we again add the constraint E to the feasible region (with another error of at most $\frac{k(m')}{\sqrt{m'}}$ added on), then this is bounded above by $2(\epsilon - 2\frac{k(m')}{\sqrt{m'}})^2$. Union bounding over i yields the result. □

3.2.4 Some Insights

We have established that, with probability at least $(1-\nu)$, the following holds:

$$\bar{\delta}(\mathbb{P}_f) - \zeta \lesssim \inf_{m' \in \mathbb{Z}^+} \sqrt{\frac{log\frac{1}{\nu} + m'|\mathcal{Y}|log(2)}{2m}} + \delta' + 2\delta' \tag{3.62}$$

where $\delta' = \frac{k(m')}{\sqrt{m'}}$ and we can usually take $\zeta \approx 0$ (as we can make this arbitrarily small with a large enough network, due to [49] and lemma 5 if we train on cross-entropy errors). $k(m')$ is trivially less than or equal to m', but it is generally going to be quite small since it is dependent on a statement only requiring the *existence of functions* satisfying an empirical deviation bound. This is in contrast to classical statistical learning theory bounds which instead require *for all functions* statements of the same sort. Furthermore, $k(m')$ is not strictly increasing with model complexity. On the contrary, $k(m')$ can decrease as the hypothesis space grows (given that we maintain $\mathcal{W}$ continuity), since having more functions will increase the probability of such existences. By Theorem 4, we can also assume that $\frac{k(m')}{m'} \to 0$ as $m' \to 0$. These intuitions tell us that the decomposition in Theorem 1 has successfully extracted a good amount of the problem's complexity into the term $I(X;Z)$. The primary complexity term in $\bar{\delta}(\mathbb{P}_f)$ - given a sufficiently complex hypothesis space - arises from the complexity of the class variable itself.

3.3 Experiments

3.3.1 How These Bounds Solve Experimental Discrepancy

We argue that the bounds presented in this dissertation explain the experimental discrepancy that we've alluded to a few times in this dissertation. These tightened, less sensitive bounds imply that, in many cases, it is simply not optimal in terms of information losses to compress a neural network's input. This can be seen visually in Figure 3.2. Here we have set up a toy classification problem with $H(Y) = log_2(10)$, $H(X) = 21$, and $I(Y;Z^*) = H(Y)\left(1 - e^{-\frac{I(X;Z^*)}{2}}\right)$. The information quantities in this toy example are thus similar to MNIST [79]. We have plotted $I(Y;Z^*)$ along with the bounds of this dissertation (assuming $\zeta \approx 0, k(m') \approx 0$) for $m = 10,000$, $5,000$, and $2,000$ data points. We see that very little to nothing can be gained by compression in the $m = 10,000$ and $m = 5,000$ cases. Serious gains can only be obtained in the $m = 2,000$ case. On the right side of this figure, we plot the old bounds, which predicts a peak at around 5 bits even for $10,000$ data points. Thus the lack of compression found experimentally on smaller datasets is explained by our new bounds, but not by the old ones.

But if the entropy of the feature space becomes large, as we've made it for the third plot in this figure, compression becomes important even with our new bounds. This helps to explain why neural networks seem to yield compression on 'harder' datasets, but do not on 'easier' ones.

3.3.2 Tightness of Bounds

For these experiments, we have used the MINE-f [9] estimator of mutual information for $I(X;Z)$ quantities. We assume that $\hat{I}(Y;\hat{Z})$ is equal to $H(Y)$, and estimate $I(Y;\hat{Z})$ via validation error probability and Fano's inequality. To make the classifier representation stochastic, we used permanent dropout with a rate of 0.7. All classifiers are trained for

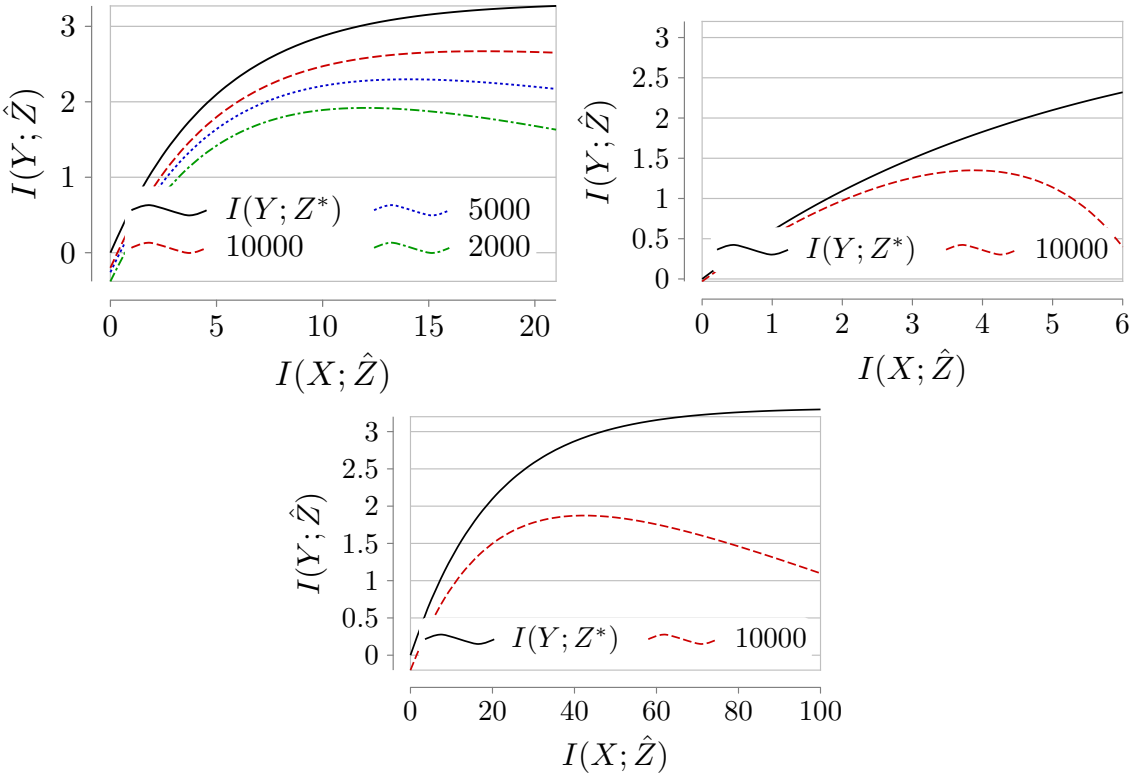

Figure 3.2: (left) New bounds on a low entropy feature space (right) Old bounds on the same space. (Bottom) New bounds on a high entropy feature space.

$10,000$ epochs, and all information estimations are performed for 2000 epochs. All neural networks are trained with the Adam optimizer. All models used a learning rate of 5×10^{-4}.

We first tested the non-asymptotic bound of Theorem 5 on four of the datasets provided by OpenML [100] across several training data sizes (dependent on the overall size of the dataset in question). Our classifier consisted of a neural network with a single hidden layer of 1000 units. The results are plotted in figure 3.3. We took a confidence interval $\nu = 0.5$ for the plot of the bound, and plotted the mean value of ten experiments for the 'true' 50% confidence interval (assuming a symmetric distribution). We estimated $k(m')$ via $k_c m'^r$ with $r < \frac{1}{2}$. In each case, we estimated k_c and r in sample for the smallest tested training data size. This, of course, only gives us a 'functional behavior' experiment, but we do see that this behavior is consistent with the true values.

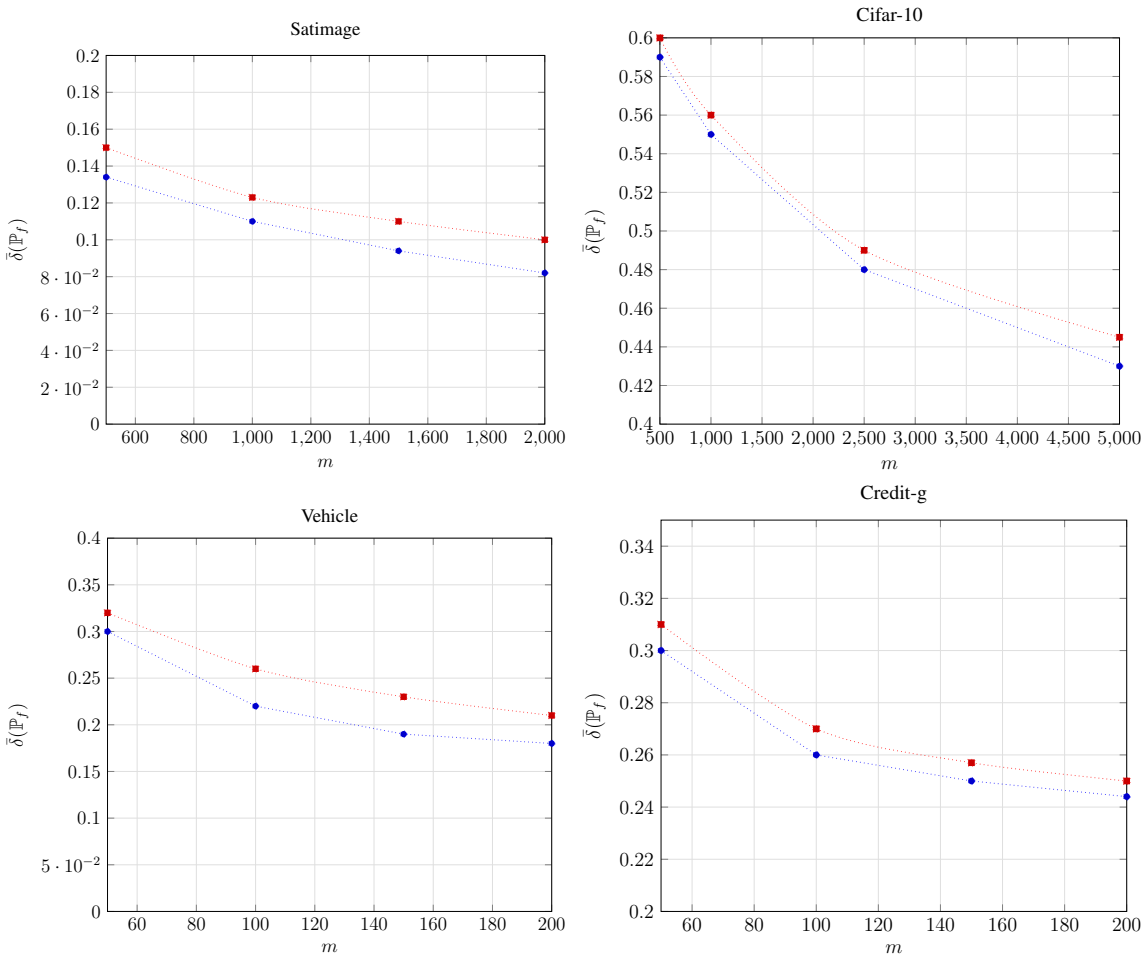

Figure 3.3: $(\bar{\delta}(\mathbb{P}_f) - \zeta)$ for several datasets. (Blue) True confidence interval, (Red) bound [Theorem 5].

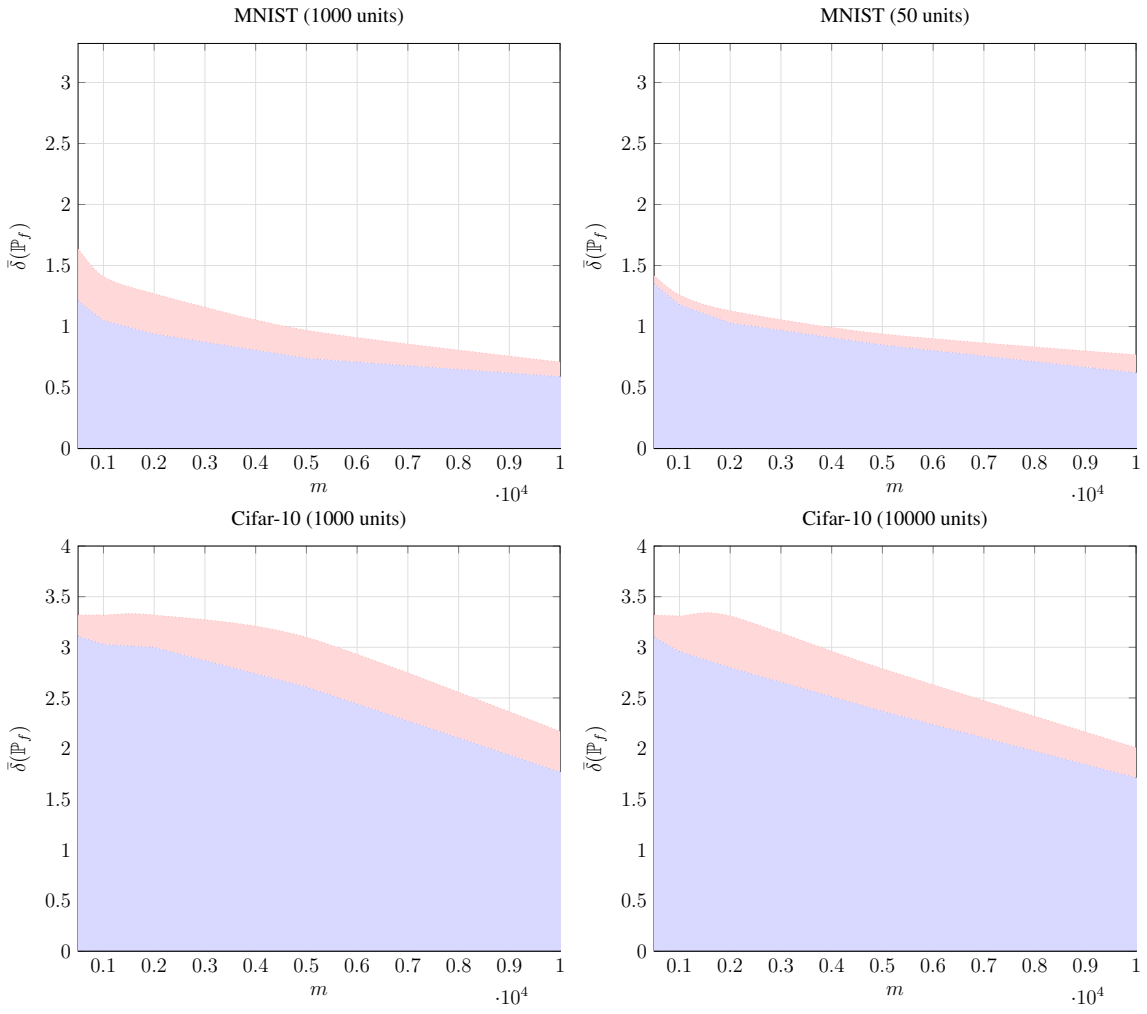

Figure 3.4: $I^{(1)}_{Loss}$ for MNIST over varying architectures. (Blue) True confidence interval, (Red) Information bound [Theorem 1].

We then tested the bound of Theorem 1 for MNIST and Cifar-10 using the true value of $\bar{\delta}(\mathbb{P}_f)$ in each case. The results are shown in Figure 3.4. Each dataset is experimented on with a classifier given by a fully connected neural network with single hidden layer, with varying hidden layer sizes. The deviations here are to show that the bound is decent across differing architectures. The bound is quite close to the true confidence interval in each case.

3.4 Chapter Conclusion

This chapter presented new bounds on information losses from finite data. This began in the form of a relationship between these losses, the expected total variation of the neural model, and the information held in the hidden representation of the feature space. Then, by bounding the total variation term without invoking any more dependence on model complexity, we obtained bounds that are much tighter and less sensitive to $I(X;Z)$ than previous theory. The chapter provided applications of this theoretical framework, focusing primarily on relevant contradictory experimental work that previously went unexplained. It concluded with experiments showing that the bound presented in this chapter corresponds well to experiment.

CHAPTER 4
THE MAXIMUM MUTUAL INFORMATION OF VARYING ARCHITECTURES

4.1 MMI Calculations

There is a strong relationship between the value of $I(X;Z)$ that we obtain from a neural architecture and the design of that architecture. Intuitively, larger structures will result in higher such mutual information values while smaller ones will result in lower values. In this chapter, we will study this relationship in a more exact manor. Doing so will allow for information-theoretic driven architecture design.

We will primarily be interested in a quantity which we will be calling the *Maximum Mutual Information*. This is essentially just the maximum value of $I(X;Z)$ that can be obtained with a given architecture.

4.2 Single Layer Linear Networks (Fully Connected and Convolutional)

4.2.1 Fully Connected Case

We begin by deriving the MMI of a linear network with a standardized Gaussian input. We consider the constrained problem in which the weight matrices W are constrained by Frobenius norm.

Theorem 6. *Let Σ_x be a positive definite matrix and let $\sigma^2 > 0$. Let N_0 and N_1 be natural numbers. Let $\mathcal{N}(\mu; A)$ denote the Gaussian distribution with mean μ and covariance matrix*

A. Let:

$$\begin{aligned} X &\sim \mathcal{N}(0; \Sigma_x), & X &\in \mathbb{R}^{N_0} \\ Z|X, W, b &\sim \mathcal{N}(WX + b;\ \sigma^2 Id_{N_1}), & Z &\in \mathbb{R}^{N_1} \end{aligned} \tag{4.1}$$

Where $W \in \mathbb{R}^{N_1 \times N_0}$.

Let $\tilde{N} = \min(N_0, N_1)$. *Let* $\Sigma_{x,\tilde{N}}$ *denote* $\tilde{N} \times \tilde{N}$ *diagonal matrix containing the* $\tilde{N}$ *largest eigenvalues of* Σ_x. *Let* $\lambda^x_{\tilde{N}}$ *denote the smallest eigenvalue of* $\Sigma_{x,\tilde{N}}$, *and let:*

$$\rho_{\tilde{N}} \triangleq \sigma^2 \left(\frac{\tilde{N}}{\lambda^x_{\tilde{N}}} - Tr(\Sigma^{-1}_{x,\tilde{N}}) \right) \tag{4.2}$$

Let:

$$I_{W,b}(X; Z) \triangleq \mathbb{E}_{\sim p(x,z|w)} \left[log \frac{p(z|x,w)}{p(z|w)} \right] \tag{4.3}$$

Let $F \geq \rho_{\tilde{N}}$, *and let:*

$$\textit{MMI}(X; Z) \triangleq \sup_{Tr(W^T W) \leq F} I_{W,b}(X; Z) \tag{4.4}$$

Then:

$$\textit{MMI}(X; Z) = \frac{\tilde{N}}{2} log \left(\frac{F + \sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}})}{\sigma^2 \tilde{N}} \right) + \frac{1}{2} log\ |\Sigma_{x,\tilde{N}}| \tag{4.5}$$

Proof. Since X is Gaussian and the network is linear, Z is Gaussian for all W, b and we

have:

$$\begin{aligned} I_{W,b}(X;Z) &= H(Z) - H(Z|X) \\ &= \frac{1}{2} log\ \frac{|\Sigma_z|}{|\Sigma_{z|x}|} \\ &= \frac{1}{2} log\ \frac{|\sigma^2 Id_{N_1} + W\Sigma_x W^T|}{|\sigma^2 Id_{N_1}|} \\ &= \frac{1}{2} log\ |Id_{N_1} + \frac{1}{\sigma^2} W\Sigma_x W^T| \end{aligned} \tag{4.6}$$

Now, by the matrix determinant lemma, we have:

$$|Id_{N_1} + \frac{1}{\sigma^2} W\Sigma_x W^T| = |\sigma^2 \Sigma_x^{-1} + W^T W| \cdot |\frac{1}{\sigma^2}\Sigma_x| \tag{4.7}$$

and so we can extract the dependence of (4.27) on W to obtain:

$$\text{MMI}(X;Z) = \frac{1}{2} log\ |\frac{1}{\sigma^2}\Sigma_x| + \frac{1}{2} sup\ log\ |Z| \tag{4.8}$$

where $Z = \sigma^2 \Sigma_x^{-1} + W^T W$. Due to the positive definiteness of Z and Hadmard's inequality, we can cast this transformed optimization problem into the realm of eigenvalues as:

$$\begin{aligned} &\sup_{\tilde{\lambda}_1, \cdots, \tilde{\lambda}_{N_0}} \sum_{i=1}^{N_0} log \left(\tilde{\lambda}_i + \frac{\sigma^2}{\lambda_i^x} \right) \\ &\text{s.t. } \sum_{i=1}^{N_0} \tilde{\lambda}_i \leq F \\ &\text{and } \tilde{\lambda}_i \geq 0,\ i = 1, 2, \cdots, N_0 \\ &\text{and } \tilde{\lambda}_i = 0,\ \text{for at least } \max(0, N_0 - N_1) \text{ values of } i \end{aligned} \tag{4.9}$$

where λ_i^x is the i^{th} largest eigenvalue of Σ_x, and $\tilde{\lambda}_i$ is the i^{th} un-ordered eigenvalue of W^TW. The final constraint comes from the fact that W^TW is only rank $\max(N_0, N_1)$.

We will now show that the mandatory 0-valued eigenvalues of W^TW, when they exist ($N_0 > N_1$), must be placed on the indices $i = N_1 + 1, \cdots, N_0$, as these correspond to the largest values of $\frac{\sigma^2}{\lambda_i^x}$. Indeed, suppose that we have placed a nonzero eigenvalue on one of these indices (say index p WLOG, and say WLOG that we set this eigenvalue to l) in such a way that all of the constraints are met. Then we must have placed a zero-valued eigenvalue on another index (which we will call q WLOG, $q < p$). Then the objective function can be increased (without violating any constraints) by taking l units of eigenvalue off of index p and placing it on index q, since:

$$\begin{aligned} exp\left(log\left(l + \frac{\sigma^2}{\lambda_p^x}\right) + log\left(\frac{\sigma^2}{\lambda_q^x}\right)\right) &= \frac{l\sigma^2}{\lambda_q^x} + \frac{\sigma^4}{\lambda_p^x\lambda_q^x} \\ &\leq \frac{l\sigma^2}{\lambda_p^x} + \frac{\sigma^4}{\lambda_p^x\lambda_q^x} \\ = exp\left(log\left(\frac{\sigma^2}{\lambda_p^x}\right) + log\left(l + \frac{\sigma^2}{\lambda_q^x}\right)\right) & \end{aligned} \tag{4.10}$$

And so this cannot be a solution to our optimization problem. We are thus left with the following problem:

$$\begin{aligned} &\sup_{\tilde{\lambda}_1, \cdots, \tilde{\lambda}_{\tilde{N}}} \sum_{i=1}^{\tilde{N}} log\left(\tilde{\lambda}_i + \frac{\sigma^2}{\lambda_i^x}\right) \\ &\text{s.t. } \sum_{i=1}^{\tilde{N}} \tilde{\lambda}_i \leq F \\ &\text{and } \tilde{\lambda}_i \geq 0,\ i = 1, 2, \cdots, \tilde{N} \end{aligned} \tag{4.11}$$

This is a classic 'water-filling' problem with heights given by scaled versions of the inverses

of the first $\tilde{N}$ eigenvalues of Σ_x. Thus, for a given 'water level', $\mu^*(F)$, a solution is readily available, being given by:

$$\tilde{\lambda}_i = \max\left(0, \mu^*(F) - \frac{\sigma^2}{\lambda_i^x}\right) \tag{4.12}$$

However, finding the relationship between $\mu^*(F)$ and F requires additional work, as $\mu^*(F)$ must adhere to the equation:

$$\sum_i \max\left(0, \mu^*(F) - \frac{\sigma^2}{\lambda_i^x}\right) = F \tag{4.13}$$

We will show that our assumption, $F \geq \rho_{\tilde{N}}$, yields a consistent solution in which all maximums of (4.12) are obtained in the second argument. To see this, note that under such a solution, (4.13) yields:

$$\mu^*(F) = \frac{F + \sigma^2 Tr(\Sigma_{x,\tilde{N}}^{-1})}{\tilde{N}} \tag{4.14}$$

Which is consistent with (4.12) because, for each i, we have the following sequence of inequalities:

$$\frac{F + \sigma^2 Tr(\Sigma_x^{-1})}{\tilde{N}} - \frac{\sigma^2}{\lambda_i^x} \geq \frac{1}{\tilde{N}}\left(F - \rho_{\tilde{N}}\right) \geq 0 \tag{4.15}$$

Thus, in all, we have an MMI of

$$\begin{aligned}
&\frac{1}{2} log|\frac{\Sigma_x}{\sigma^2}| + \frac{\tilde{N}}{2} log\left(\frac{F + \sigma^2 Tr(\Sigma_{x,\tilde{N}}^{-1})}{\tilde{N}}\right) \\
&\qquad + \frac{1}{2}\sum_{i=\tilde{N}+1}^{N_0} log(\frac{\sigma^2}{\lambda_i^x}) \\
&= \frac{1}{2}\sum_{i=1}^{\tilde{N}} log(\frac{\lambda_i^x}{\sigma^2}) + \frac{\tilde{N}}{2} log\left(\frac{F + \sigma^2 Tr(\Sigma_{x,\tilde{N}}^{-1})}{\tilde{N}}\right) \\
&= \frac{1}{2} log|\Sigma_{x,\tilde{N}}| + \frac{\tilde{N}}{2} log\left(\frac{F + \sigma^2 Tr(\Sigma_{x,\tilde{N}}^{-1})}{\sigma^2 \tilde{N}}\right)
\end{aligned} \tag{4.16}$$

□

Lemma 8. *Take all of the assumptions from Theorem 6 except for the assumption that $F \geq \rho$. Let e_k denote the eigenvector of Σ_x corresponding to the k^{th} smallest eigenvalue, respectively denoted λ_k^x. Let K be a natural number, $K < \tilde{N}$, and let $\Sigma_{x,\tilde{N}-K}$ denote the $(\tilde{N}-K) \times (\tilde{N}-K)$ diagonal matrix containing the $\tilde{N}-K$ largest eigenvalues of Σ_x. Now, let:*

$$\rho_{\tilde{N}-K} \triangleq \sigma^2 \left(\frac{\tilde{N}-K}{\lambda_{\tilde{N}-K}^x} - Tr\left(\Sigma_{x,\tilde{N}-K}^{-1}\right) \right), \tag{4.17}$$

Then we have:

$$0 = \rho_1 \leq \cdots \leq \rho_{\tilde{N}-K} \leq \rho_{\tilde{N}-K+1} \leq \cdots \leq \rho_{\tilde{N}-1} \leq \rho_{\tilde{N}} \tag{4.18}$$

Proof. First, ρ_1 is zero since:

$$\rho_1 = \sigma^2 \left(\frac{1}{\lambda_1^x} - \frac{1}{\lambda_1^x} \right) \tag{4.19}$$

Next, we note that the difference $\rho_{\tilde{N}-K+1} - \rho_{\tilde{N}-K}$ is given by:

$$\begin{aligned}
&\sigma^2 \left(\frac{\tilde{N}-K+1}{\lambda_{\tilde{N}-K+1}^x} - \frac{\tilde{N}-K}{\lambda_{\tilde{N}-K}^x} - \frac{1}{\lambda_{\tilde{N}-K+1}^x} \right) \\
&= \sigma^2 \left(\frac{\tilde{N}-K}{\lambda_{\tilde{N}-K+1}^x} - \frac{\tilde{N}-K}{\lambda_{\tilde{N}-K}^x} \right) \\
&\geq \sigma^2 \left(\frac{\tilde{N}-K}{\lambda_{\tilde{N}-K}^x} - \frac{\tilde{N}-K}{\lambda_{\tilde{N}-K}^x} \right) \\
&= 0
\end{aligned} \tag{4.20}$$

completing the proof. □

Theorem 7. *Take all of the assumptions from Theorem 6 except for the assumption that* $F \geq \rho_{\tilde{N}}$ *and take all definitions from lemma 8. Let* $\rho_{\tilde{N}-K+1} \geq F \geq \rho_{\tilde{N}-K}$. *Then* ***MMI***$(X;Z)$ *is given by:*

$$\frac{\tilde{N}-K}{2} log\left(\frac{F+\sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K})}{\sigma^2(\tilde{N}-K)}\right) + \frac{1}{2} log\,|\Sigma_{x,\tilde{N}-K}| \tag{4.21}$$

Proof. We can follow the proof of Theorem 6 up until the optimization problem given by (4.11), whose solution will now be different. Again, we need to find a solution consistent with (4.12) and (4.13). We claim that our assumption, $\rho_{\tilde{N}-K+1} > F \geq \rho_{\tilde{N}-K}$ yields a consistent solution in which the $\tilde{\lambda}_i$ are zero for $i > \tilde{N}-K$ and nonzero otherwise. Under such a solution, (4.13) yields $\mu^*(F) = \frac{F+\sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K})}{\tilde{N}-K}$. We will show that this is consistent with (4.12). First, define $l(i) \triangleq \frac{F+\sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K})}{\tilde{N}-K} - \frac{\sigma^2}{\lambda^x_i}$. Then, if $i \leq \tilde{N}-K$, we have the following inequalities on $l(i)$ ultimately showing the positivity of this value, and therefore the consistency of the solution:

$$l(i) \geq \frac{F+\sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K})}{\tilde{N}-K} - \frac{\sigma^2}{\lambda^x_{\tilde{N}-K}} = \frac{1}{\tilde{N}-K}\left(F-\rho_{\tilde{N}-K}\right) \geq 0 \tag{4.22}$$

On the other hand, if $i > \tilde{N}-K$, then:

$$\begin{aligned} l(i) &\leq \frac{F+\sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K})}{\tilde{N}-K} - \frac{\sigma^2}{\lambda^x_{\tilde{N}-K+1}} \\ &= \frac{F+\sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K+1}) - \frac{\sigma^2}{\lambda^x_{\tilde{N}-K+1}}}{\tilde{N}-K} - \frac{\sigma^2}{\lambda^x_{\tilde{N}-K+1}} \\ &= \frac{1}{\tilde{N}-K}\left(F-\rho_{K+1}\right) < 0 \end{aligned} \tag{4.23}$$

Under this solution, the objective function value is given by:

$$\sum_{i=\tilde{N}-K+1}^{\tilde{N}} log\left(\frac{\sigma^2}{\lambda_i}\right) + (\tilde{N}-K)log\left(\frac{F+\sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K})}{\tilde{N}-K}\right) \tag{4.24}$$

Thus, in all, we have that MMI$(X;Z)$ is given by:

$$\frac{\tilde{N}-K}{2}log\left(\frac{F+\sigma^2 Tr(\Sigma^{-1}_{x,\tilde{N}-K})}{\sigma^2(\tilde{N}-K)}\right) + \frac{1}{2}log\left(\frac{|\Sigma_{x,\tilde{N}}|}{\prod_{i=\tilde{N}-K+1}^{\tilde{N}}\lambda_i}\right)$$

The only factors remaining in post-cancellation of the second term are the eigenvalues of $\Sigma_{x,\tilde{N}}$ with indices smaller than or equal to $\tilde{N}-K$. This completes the proof. □

4.2.2 Convolutional Case

Theorem 8. *Let Σ_x be a positive definite matrix and let $\sigma^2 > 0$. Let N_0, N_B, and N_f be natural numbers such that N_B divides N_0, Let:*

$$\begin{aligned} X &\sim \mathcal{N}(0;\Sigma_x), & X &\in \mathbb{R}^{N_0} \\ Z|X,W,b &\sim \mathcal{N}(X \circledast W + b;\ \sigma^2 Id_{N_1}), & Z &\in \mathbb{R}^{N_0/N_B} \end{aligned} \tag{4.25}$$

Where $W \in \mathbb{R}^{N_f \times N_b}$ and:

$$X \circledast W = \begin{bmatrix} W\tilde{X}_1 \\ W\tilde{X}_2 \\ \cdots \\ W\tilde{X}_{N_0/N_B} \end{bmatrix} \tag{4.26}$$

with $\tilde{X}_j$ denoting the slice of X on indices $(j-1)N_1+1$ through jN_1. Thus $X \circledast W$ is a convolution applied to a vectorized input with non-overlapping stride and N_f filters.

Suppose Σ_x is block diagonal with identical blocks given by the $N_B \times N_B$ matrix $\Sigma_{\tilde{x}}$. Let $M_{LFC}(F; \Sigma_{\tilde{x}}, N_b, N_f)$ denote the maximum mutual information of the linear fully connected network given by Theorems 6 and 7 when the input covariance matrix is given by $\Sigma_{\tilde{x}}$, and with N_0 in that theorem set to N_B, and with N_1 in that theorem set to N_f. Let:

$$\textit{MMI}(X;Z) \triangleq \sup_{Tr(W^T W) \leq F} I_{W,b}(X;Z) \tag{4.27}$$

Then:

$$\textit{MMI}(X;Z) = \frac{N_0}{N_B} M_{LFC}(F; \Sigma_{\tilde{x}}, N_b, N_f) \tag{4.28}$$

Proof. We can view the output of the convolution as the following matrix product:

$$X \circledast W = \begin{bmatrix} W & 0 & \cdots & 0 \\ 0 & W & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W \end{bmatrix} X \tag{4.29}$$

Hence we can denote this block diagonal matrix as $\tilde{W}$ and follow the proof of Theorem 6 until equation (4.7) with $\tilde{W}$ in place of W. From here, we can further factor as follows:

$$\begin{aligned} log\ |\sigma^2 \Sigma_x^{-1} + W^T W| &= \frac{N_0}{N_B} log\ |\sigma^2 \Sigma_{\tilde{x}}^{-1} + W^T W| \\ log\ \left|\frac{\Sigma_x}{\sigma^2}\right| &= \frac{N_0}{N_B} log\ \left|\frac{\Sigma_{\tilde{x}}}{\sigma^2}\right| \end{aligned} \tag{4.30}$$

from which we are back to the original optimization problem in the proof of Theorems 6 and 7, just multiplied by a factor of N_1. □

4.3 Single Layer Relu Networks

4.3.1 Relu Activations

Lemma 9. *Let $Y \in \mathbb{R}^d$. Then:*

$$relu(Y)relu(Y)^T \preceq YY^T \tag{4.31}$$

$$relu(Y)relu(Y)^T \preceq relu(YY^T) \tag{4.32}$$

And, in general, YY^T and $relu(YY^T)$ are incomparable in the Loewner order. Finally, let $c \in \mathbb{R}^d$ be a vector with non-negative elements. Then:

$$(relu(Y) - c)(relu(Y) - c)^T \preceq (Y - c)(Y - c)^T \tag{4.33}$$

Proof. For notational convenience, we will denote the relu function as r throughout this proof. Without loss of generality, we suppose that all negative components of Y are contained in the last K indices. That is, $y_i \geq 0$ when $i = 1, \cdots, d - K$, and $y_i < 0$ when $i = d - K + 1, \cdots, d$.

We begin by proving the first inequality. Note that both YY^T and $r(Y)r(Y)^T$ are rank-1. The single nonzero eigenvalue of YY^T is $\|Y\|_2^2$ and the single nonzero eigenvalue of $r(Y)r(Y)^T$ is $\|r(Y)\|^2$. Clearly, the former is larger than the later.

We now move on to proving the second inequality. Under our WLOG assumption at the beginning of this proof, $r(Y)r(Y)^T$ is a rank-1 matrix with zeros outside of the top-left block of size $(d - K) \times (d - K)$. Denote this block matrix as A.

Now, $r(YY^T)$ will also contain A in its top-left block as all elements in the set

$$\{y_i y_j : \ i, j \leq d - K\} \tag{4.34}$$

are non-negative. Similarly, all off-diagonal elements outside of this block will be zero because all elements in the set $\{y_i y_j, i \leq d-K, j > d-K \vee i > d-K, j \leq d-K\}$ are negative. Finally, all diagonal elements outside of this block are positive, taking values from the set $\{y_i^2, i > d-K\}$. We can thus write:

$$r(YY^T) = r(Y)r(Y)^T + D \tag{4.35}$$

where D is a diagonal matrix with 0 along the top-left $d-K$ diagonal elements, and with positive elements along the bottom-right K diagonal elements. As such, D is positive semidefinite, which concludes the proof of the second inequality.

To see that the two right-hand-side quantities are generally incomparable, consider the vector $Y = \begin{bmatrix} 1 & -1 \end{bmatrix}^T$.

Finally, to prove the last inequality, we note again that both sides of the inequality are rank-1 matrices whose single nonzero eigenvalues are given by the squared norms of their corresponding input vectors. We thus only need to show that

$$\|r(Y) - c\|^2 \leq \|Y - c\|^2 \tag{4.36}$$

To that end, we can write the difference between the RHS and the LHS as:

$$\|Y\|^2 - \|r(Y)\|^2 + 2c^T(r(Y) - Y) \tag{4.37}$$

Now, $r(Y) - Y$ and c both contain only non-negative elements, so the third term is also non-negative. Furthermore, $\|Y\|^2 \geq \|r(Y)\|^2$, so (4.37) is non-negative. This concludes the proof of the final inequality. □

Lemma 10. *Take all of the assumptions of either Theorem 6, Theorem 7, or Theorem 8. Let W, b be fixed and define Z_{relu} through a new model given by:*

$$Z_{relu}|X, W, b \sim \mathcal{N}(\textit{relu}(WX + b);\ \sigma^2 Id_{N_1}) \tag{4.38}$$

for the fully connected case, or, for the convolutional case:

$$Z_{relu}|X, W, b \sim \mathcal{N}(\textit{relu}(X \circledast W + b);\ \sigma^2 Id_{N_1}) \tag{4.39}$$

Let Σ_Z denote the covariance matrix of Z as defined in Theorem 6 and let $\Sigma_{Z_{relu}}$ denote the covariance matrix of Z_{relu}. Then:

$$\Sigma_{Z_{relu}} \preceq \Sigma_Z \tag{4.40}$$

Proof. Let η denote a multivariate Gaussian with covariance $\sigma^2 Id_{N_1}$. We further denote $S \triangleq WX + b$. Then

$$\begin{aligned} Z &= S + \eta \\ Z_{relu} &= relu(S) + \eta \end{aligned} \tag{4.41}$$

Let μ_r denote the expectation of $relu(S)$. We thus obtain:

$$\begin{aligned} \Sigma_{Z_{relu}} &= \mathbb{E}\left[Z_{relu} Z_{relu}^T\right] - \mu_r \mu_r^T \\ &= \sigma^2 Id_{N_1} + \mathbb{E}\left[relu(S) relu(S)^T\right] - \mu_r \mu_r^T \\ &\preceq \sigma^2 Id_{N_1} + \mathbb{E}\left[SS^T\right] - \mu_r \mu_r^T \\ &= \Sigma_Z - \mu_r \mu_r^T \end{aligned} \tag{4.42}$$

where the inequality follows from lemma 9. The proof follows immediately from noting that $\mu_r \mu_r^T$ is positive semidefinite.

For the convolutional case, replace W in the above proof with $\tilde{W}$ from the proof of Theorem 8. □

Lemma 11. *Take all of the assumptions of either Theorem 6 or Theorem 7 and all definitions from lemma 10. Denote the marginal probability laws of Z and Z_{relu} as $\mathbb{P}$ and $\hat{\mathbb{P}}$ with densities denoted $p(z)$ and $\hat{p}(z)$. Let $\mathcal{B}(Z)$ be the set of Borel measurable sets on $\mathcal{Z}$, and let δ denote the total variation distance between $\mathbb{P}$ and $\hat{\mathbb{P}}$:*

$$\delta = \sup_{A \in \mathcal{B}(Z)} |\mathbb{P}(A) - \hat{\mathbb{P}}(A)| = \frac{1}{2} \int |p(z) - \hat{p}(z)| dz \tag{4.43}$$

Finally, let $h_2(\cdot)$ denote the binary entropy function. Then, for all W, b:

$$|I_{W,b}(X;Z) - I_{W,b}(X;Z_{relu})| \leq \delta log((\frac{2\pi e}{\delta})^{N_1} |\Sigma_z|) + 2h_2(\delta) \tag{4.44}$$

Proof. Let γ denote the standard maximal coupling between $\mathbb{P}$ and $\hat{\mathbb{P}}$ on the variables $(\tilde{Z}, \hat{Z})$ (not to be confused with the *conditional maximal coupling* defined earlier). Let the interim variables defining γ be denoted as J, U, V and W as they are in construction 1 of this dissertation. Then, as $H(\tilde{Z}|X) = H(\hat{Z}|X)$, we have:

$$|I(X;Z) - I(X;\hat{Z})| = |H(\tilde{Z}) - H(\hat{Z})| \tag{4.45}$$

We can decompose these terms as:

$$\begin{aligned} H(\tilde{Z}) &= H(\tilde{Z}|J) + H(J) - H(J|\tilde{Z}) \\ H(\hat{Z}) &= H(\hat{Z}|J) + H(J) - H(J|\hat{Z}) \end{aligned} \tag{4.46}$$

The $H(J)$ terms cancel in the subtraction. Furthermore, both $H(J|\tilde{Z})$ and $H(J|\hat{Z})$ are bounded by $H(J) = h_2(\delta)$. Thus, by an application of the triangle inequality, we have:

$$|I(X;Z) - I(X;\hat{Z})| \leq |H(\tilde{Z}|J) - H(\hat{Z}|J)| + 2h_2(\delta) \tag{4.47}$$

We can further decompose the remaining terms as:

$$\begin{aligned} H(\tilde{Z}|J) &= (1-\delta)H(U) + \delta H(V) \\ H(\tilde{Z}|J) &= (1-\delta)H(U) + \delta H(W) \end{aligned} \tag{4.48}$$

leaving us with:

$$|I(X;Z) - I(X;\hat{Z})| \leq \delta|H(V) - H(W)| + 2h_2(\delta) \tag{4.49}$$

Next we will prove the following variance inequalities:

$$\Sigma_V, \Sigma_W \preceq \frac{\Sigma_Z}{\delta} \tag{4.50}$$

where Σ_V and Σ_W are the covariance matrices of V and W. From these inequalities, we have by entropy maximization that:

$$H(V), H(W) \leq \frac{1}{2} log \left((\frac{2\pi e}{\delta})^{N_1} |\Sigma_Z| \right) \tag{4.51}$$

and the conclusion of the theorem immediately holds. To show that the above variance

inequalities hold, we first explicitly write densities of V and W in the maximal coupling:

$$\begin{aligned} p_V(z) &= \frac{p(z) - \max(p(z), \hat{p}(z))}{\delta} \\ p_W(z) &= \frac{\hat{p}(z) - \max(p(z), \hat{p}(z))}{\delta} \end{aligned} \tag{4.52}$$

From which we can immediately see that:

$$\begin{aligned} \Sigma_V = &\frac{1}{\delta} \int (z - \mu_V)(z - \mu_V)^T p(z) \delta dz \\ &- \frac{1}{\delta} \int (z - \mu_V)(z - \mu_V)^T \max(p(z), \hat{p}(z)) dz \\ &= \frac{\Sigma_Z}{\delta} - \frac{A}{\delta} \preceq \frac{\Sigma_Z}{\delta} \end{aligned} \tag{4.53}$$

where A, meant as a placeholder for the second integral in (4.53), is a positive semidefinite matrix. We can similarly derive that:

$$\Sigma_W \preceq \frac{\Sigma_{Z_{relu}}}{\delta} \preceq \frac{\Sigma_Z}{\delta} \tag{4.54}$$

where the second inequality follows from lemma 10, completing the proof. □

Lemma 12. *Take all of the assumptions of either Theorem 6 or Theorem 7, and take all definitions from lemmas 10 and 11. Then*

$$H(Z_{relu}) \leq H(Z) \tag{4.55}$$

Proof. Since Z_{relu} has covariance matrix $\Sigma_{Z_{relu}}$, we know from the entropy maximization

principal of multivariate Gaussians that:

$$\begin{aligned} H(Z_{relu}) &\leq \frac{1}{2} log\left((2\pi e)^{N_1}|\Sigma_{Z_{relu}}|\right) \\ &\leq \frac{1}{2} log\left((2\pi e)^{N_1}|\Sigma_Z|\right) = H(Z) \end{aligned} \tag{4.56}$$

where the second inequality follows from lemma 10 and the monotanacity of the determinant function. □

Theorem 9. *Take all of the assumptions of either Theorem 6 or Theorem 7, and take all definitions from lemmas 10 and 11. Then the results of Theorem 6 and Theorem 7 hold for* ***MMI***$(X; Z_{relu})$*. That is,*

$$\textbf{\textit{MMI}}(X; Z_{relu}) = \textbf{\textit{MMI}}(X; Z) \tag{4.57}$$

Proof. First, we note that, for all W, b, we have:

$$\begin{aligned} I_{W,b}(X; Z_{relu}) &= H_{W,b}(Z_{relu}) - H(Z_{relu}|X) \\ &= H_{W,b}(Z_{relu}) - H(Z|X) \\ &\leq H_{W,b}(Z) - H(Z|X) = I_{W,b}(X : Z) \end{aligned} \tag{4.58}$$

where the inequality follows from lemma 12. It follows immediately that:

$$\text{MMI}(X; Z_{relu}) \leq \text{MMI}(X; Z) \tag{4.59}$$

We now show that MMI$(X; Z)$ is an achievable value for $I_{W,b}(X; Z_{relu})$ given the constraint $Tr(W^T W) \leq F$. Fix $\epsilon > 0$. Then given any value of F, we can set b large enough in each dimension such that $\hat{\mathbb{P}}(\cup_{i=1}^{N_1}\{z|z_i < 0\})$ is less than ϵ for all W satisfying $Tr(W^T W) \leq F$. When this is the case, the total variation between $\mathbb{P}$ and $\hat{\mathbb{P}}$ is also bounded

above by ϵ. Then by lemma 11, there exists $b^* \in \mathbb{R}^{N_1}$ such that:

$$|I_{W,b^*}(X;Z) - I_{W,b^*}(X;Z_{relu})| \leq \epsilon log\left(\frac{a}{\epsilon^{N_1}}|\Sigma_Z|\right) + h_2(\epsilon) \tag{4.60}$$

for all W satisfying this constraint. The right hand side of (4.60) is a continuous function of ϵ, which we will denote as $g(\epsilon)$, over $(0,\infty)$, which staisfies:

$$\lim_{\epsilon \to 0^+} g(\epsilon) = 0 \tag{4.61}$$

Thus we can achieve $I_{W,b^*}(X;Z_{relu}) \geq I_{W,b^*}(X;Z) - g(\epsilon)$ for all W satisfying the constraint. Since $g > 0$ can be made arbitrary small via (4.61), we can achieve:

$$I_{W,b^*}(X;Z_{relu}) \geq I_{W,b^*}(X;Z) \; \forall \; W \text{ s.t. } Tr(W^TW) \leq F \tag{4.62}$$

Inputting the MMI achieving matrix from Theorems 6 and 7 into (4.62) yields the result. □

4.4 Multilayer Networks

Theorem 10. *Take all assumptions and definitions from the previous theorems corresponding to a fully connected network, but assume that we are using a K layer neural network instead of a single layer network, with the noise placed on the K^{th} layer only. Let $N_0, N_1, \cdots, N_K$ denote the number of hidden units in each layer. Redefine $\tilde{N}$ to $\tilde{N} \triangleq \min(N_0, N_1, \cdots, N_K)$. Then the results of those previous theorems hold.*

Proof. We can take the proof of theorem 6 by replacing W with $W_K \cdots W_2 W_1$. We will only need to note that the corresponding inner-product matrix, $W_1^T W_2^T \cdots W_K^T W_k \cdots W_2 W_1$ has rank $\tilde{N}$ (as redefined in this theorem's hypothesis). □

CHAPTER 5

INFORMATION LOSSES IN TRAINING DATA SELECTION STRATEGIES

5.1 Facility Location Optimization Mitigates Information Losses - First Approach

We will first show that an existing sampling method reduces information losses in two ways. This method is known as the *facility location function selection method* [28, 85]. The method is not new. But with these derivation, it obtains a novel information theoretic interpretation.

5.1.1 Metric Facility Location

We begin with a definition:

Definition 5. *The pseudometric* $\Delta_{\mathbb{P}_{Y|X}}(x, x')$ *is given by:*

$$\Delta_{\mathbb{P}_{Y|X}}(x, x') = \frac{1}{2} \sum_{y} |p(y|x) - p(y|x')| \tag{5.1}$$

This pseudometric will be used in a comparative sense to metrics that we can learn [18, 51, 105]. The specific details of whichever metric we end up choosing is not entirely important. We will only need it to satisfy two properties. Denoting the chosen metric as Δ, these properties are listed in Table 5.1.

For the first assumption, we note that we can always arbitrarily contract a given metric. To do so, we only need to apply a sublinear monotonic function g from the non-negative reals to $[0, 1]$ with $g(0) = 0$ to the metric. The second assumption enforces that our learned metric work best on a *local* scale. This is a common assumption. It is why neighbors estimates and patched local metric spaces are often used in manifold learning [26, 65, 80,

Property	
Under-approximant	$\Delta_{\mathbb{P}_{Y\|X}} = \Delta(x, x') + \epsilon(x, x')$
	$\epsilon(x, x') \geq 0$
Local-approximant	If $f^\flat, f^\sharp : \mathcal{X} \to \mathcal{X}$
	are measurable functions such that
	$\mathbb{E}\left[\Delta(X, f^\flat)\right] \leq \mathbb{E}\left[\Delta(X, f^\sharp)\right]$, then
	$\mathbb{E}\left[\epsilon(X, f^\flat)\right] \leq \mathbb{E}\left[\epsilon(X, f^\sharp)\right]$

Table 5.1: Metric Assumptions

95]. Before continuing, we need another definition.

Definition 6. *Let S be a training dataset. Then $\eta_{S,\Delta} : \mathcal{X} \to S$ is the assignment function that maps x to its nearest neighbor in the training dataset under Δ. Then we denote as $\delta_{p^{S,\Delta}}$ the conditional total variation for model which uses $p(y|\eta_{S,\Delta}(x))$ as its estimated conditional distribution.*

From which we will take one more assumption

Assumptions 1.

- *$p(y|\eta_{S,\Delta}(x))$ is a worse conditional distribution than that obtained by the machine learning method under study. That is,*

$$\delta_{\hat{p}} \leq \delta_{p^{S,\Delta}} \tag{5.2}$$

Next, we define the facility location function:

Definition 7. *Let $\mathcal{D}$ again denote the set of potential training data points, and let $S \subseteq \mathcal{D}$. Then the facility location function under Δ, denoted $z_\Delta : 2^{\mathcal{D}} \to \mathbb{R}$, is given by*

$$z_\Delta(S) = \mathbb{E}_{\mathbb{P}_X}\left[\Delta(X, \eta_{S,\Delta})\right] \tag{5.3}$$

Definition 8. *Let $\mathcal{D}$ denote the set of all potential training data points. Then we denote as $\mathcal{Z}$ the toset whose objects are subsets of $\mathcal{D}$ ordered by facility location function value.*

Definition 9. *Let $\mathcal{D}$ denote the set of all potential training data points. Then we denote as $\mathcal{V}$ the toset whose objects are conditional probability distributions of the form $p(y|\eta_{\Delta,S}(x))$ ordered by conditional total variation.*

We can now describe the interpretation of the facility location function selection method. First, by assumption 1, we have $\delta_{\hat{p}} \leq \delta_{p^{S,\Delta}}$. Any process which reduces $\delta_{p^{S,\Delta}}$ will thus reduce this upper bound. This hedges the risk of information losses. The next lemma shows that the facility location function selection method reduces $\delta_{p^{S,\Delta}}$.

Lemma 13. *The map $\mathcal{F} : \mathcal{Z} \to \mathcal{V}$ which sends A to $p(y|\eta_{A,\Delta}(x))$ is monotonic.*

Proof. Suppose $A \leq B$ in $\mathcal{Z}$. Then

$$\begin{aligned}
\delta\left(\mathcal{F}A\right) &= \int \frac{1}{2}\sum_{y} |p(y|x) - p(y|\eta_{\Delta}^{A}(x))| d\mathbb{P}_X \\
&= \int \Delta_{\mathbb{P}_{Y|X}}(x, \eta_{\Delta}^{A}(x)) d\mathbb{P}_X \\
&= \int \Delta(x, \eta_{\Delta}^{A}(x)) d\mathbb{P}_X + \int \epsilon(x, \eta_{\Delta}^{A}(x)) d\mathbb{P}_X \\
&= z(A) + \int \epsilon(x, \eta_{\Delta}^{A}(x)) d\mathbb{P}_X \\
&\leq z(B) + \int \epsilon(x, \eta_{\Delta}^{B}(x)) d\mathbb{P}_X \\
&= \int \Delta(x, \eta_{\Delta}^{B}(x)) d\mathbb{P}_X + \int \epsilon(x, \eta_{\Delta}^{B}(x)) d\mathbb{P}_X \\
&= \int \Delta_{\mathbb{P}_{Y|X}}(x, \eta_{\Delta}^{B}(x)) d\mathbb{P}_X \\
&= \delta\left(\mathcal{F}B\right) \qquad (5.4)
\end{aligned}$$

□

5.2 Facility Location Optimization Mitigates Information Losses - Second Approach

The goal of this section is to show that information losses are easy to deal with and lead to intuitive proofs.

For a general training data selection strategy, we emphasize the goal of finding a naive 'test' estimator which is somewhat natural to the strategy. We can then bound the conditional total variation of the 'test' estimator relatively easily. This task will often reduce to plain analysis due to the simplicity of the conditional total variation term. Doing this will often give us insight into when a given strategy is useful.

For our example, we take the selection strategy which attempts to minimize the following function of the training dataset S, $Z(S) = \mathbb{E}_{\mathcal{P}_X}[\|x - x_i\|]$ where $x_i$is the nearest neighbor of x in S. This method is known as the *facility location function selection method* [28, 85], and it is a practical, intuitive, all-at-once data selection technique. Essentially, the goal of this strategy is to pick data points such that, on average, every data point is geometrically close to some training point.

To analyze this strategy, we will use a 'test' estimator which takes into account local information near the training data. This is somewhat natural for a method which acts to reduce distances to the training data. Most importantly, we obtain some insights from this analysis. For example, we see that the theorem is asymptotic, which may imply that this method works best when we are taking somewhat large (but still limited) datasets instead of extremely small ones. Furthermore, since the Lipschitz coefficient L takes a prominent role, we may expect this method to be most effective when dealing with functions that vary rapidly (making the marginal improvement obtained by decreasing the facility location value larger). Finally, the proof of Theorem 11 uses a step in which the supremum of the gradient of $p(y|x)$ is used to replace the gradients at the training data values. This step is suitable when we do not have any apriori information on the gradients of $p(y|x)$, but may

be unsuitable otherwise. For example, if we assume that our dataset follows a manifold assumption, and critically, that $p(y|x)$ changes in regions of low density (in which we obtain gradient information just by looking at the distribution over X), then we may prefer to analyze a method that adheres more closely to Corollary 2 than to Theorem 11.

Theorem 11. *Let $\mathcal{X}$ be a bounded subset of $\mathbb{R}^d$. Suppose that we have a Lipschitz-continuous, differentiable conditional probability function $p(y|x) : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{Y}|}$ with Lipschitz coefficient L (maximized over each class variable). Let $\mathcal{S}$ denote a training dataset. Let $\mathcal{R}_i = \{x \in \mathbb{R}^d \ : \ argmin_{x' \in S} d(x, x') = x_i\}$ and consider the following 'neighbors' estimator of $p(y|x)$: $\hat{p}_{nn}(y|x) = p(y|x_i),\ x \in \mathcal{R}_i$. Finally, suppose that the machine learning algorithm of interest performs better (in terms of total variation) than $\hat{p}_{nn}(y|x)$. Then $\lim_{Z(S)\to 0} \frac{\delta_{TV}(p,\hat{p})}{Z(S)} \leq \frac{L|\mathcal{Y}|}{2}$. Or rather, $\delta_{TV}(p, \hat{p})$ is bounded above by a function which asymptotically behaves as $\frac{1}{2}L|\mathcal{Y}|Z(S)$.*

Proof. We can linearly approximate $p(y|x)$ in each region $\mathcal{R}_i$. The absolute error between $p(y|x)$ and $\hat{p}_{nn}(y|x)$ in this region is then given by:

$$|p(y|x) - \hat{p}_{nn}(y|x)| = \left|\nabla p(y|x_i)^T(x - x_i) + o(\|x - x_i\|)\right|, \forall y \in \mathcal{Y} \tag{5.5}$$

We can then bound the expected conditional total variation between p and $\hat{p}_{nn}$ via:

$$\begin{aligned}
\delta_{TV}(p, \hat{p}_{nn}) &= \frac{1}{2}\sum_y \sum_i \int_{\mathcal{R}_i} |\nabla p(y|x_i)^T(x - x_i) + o(\|x - x_i\|)| d\mathcal{P}_X \\
&\leq \frac{1}{2}\sum_y \sum_i \{\|\nabla p(y|x_i)\| \int_{\mathcal{R}_i} \|x - x_i\| d\mathcal{P}_X + \int_{\mathcal{R}_i} |o(\|x - x_i\|)| d\mathcal{P}_X\} \\
&\leq \frac{L|\mathcal{Y}|}{2} Z(S) + \frac{1}{2}|\mathcal{Y}| \sum_i \int_{\mathcal{R}_i} |o(\|x - x_i\|)| d\mathcal{P}_X
\end{aligned} \tag{5.6}$$

Denote $\eta_S : \mathbb{R}^d \to \mathbb{R}^d$ as the function which takes x to its nearest neighbor in S. Then:

$$\begin{aligned}\frac{\delta_{TV}(p, \hat{p}_{nn})}{Z(S)} &\leq \frac{L|\mathcal{Y}|}{2} + \frac{|\mathcal{Y}|}{2} \frac{\int_{\mathbb{R}^d} |o(\|x - \eta_S(x)\|)| d\mathcal{P}_X}{\int_{\mathbb{R}^d} \|x - \eta_S(x)\| d\mathcal{P}_X} \\ &\leq \frac{L|\mathcal{Y}|}{2} + \frac{|\mathcal{Y}|}{2} \int_{\mathbb{R}^d} \frac{|o(\|x - \eta_S(x)\|)|}{\|x - \eta_S(x)\|} d\mathcal{P}_X \end{aligned} \tag{5.7}$$

(For the last inequality, let $X = \frac{|o(\|x-\eta_S(x)\|)|}{\|x-\eta_S(x)\|}$ and $Y = \|x - \eta_S(x)\|$ in the Cauchy-Schwartz inequality). Now, since $\|x - \eta_S(x)\| > 0$, $Z(S) \to 0$ implies $\|x - \eta_S(x)\| \to 0$ on all but a set of measure zero (this follows from the bounded convergence theorem). Thus $\frac{|o(\|x-\eta_S(x)\|)|}{\|x-\eta_S(x)\|} \to 0$ almost surely, completing the proof. □

Corollary 2. *Take all of the assumptions of Theorem 11, but remove the assumption that* $p(y|x)$ *is Lipschitz-continuous. Let* $\tilde{Z}(S) = \sum_y \sum_{\mathcal{R}_i} \|\nabla p(y|x_i)\| \cdot \mathbb{E}_{\mathcal{P}_X} [1_{x \in \mathcal{R}_i} \cdot \|x - x_i\|]$. *Then* $\lim_{\tilde{Z}(S) \to 0} \frac{\delta_{TV}(p, \hat{p})}{\tilde{Z}(S)} \leq \frac{1}{2}$.

5.3 A Data Dependant Bound for Information Losses

We will now present a new bound on $\hat{\delta}$ which depends on the selected training set. This bound can be used for the evaluation of any active learning method.

We begin by assuming that we have some continuous, symmetric, positive definite kernel function $k(\cdot, \cdot)$ with a corresponding Reproducing Kernel Hilbert space (RKHS) $\mathcal{H}$ [11]. We will further assume that $k(x, x) = 1$ for all $x \in \mathcal{X}$ and that $k(x, x') \leq 1$ for all $(x, x') \in \mathcal{X} \times \mathcal{X}$.

Let μ be a measure. We will occasionally refer the following integral operator: $T : L^2_\mu \to \mathcal{H}$ given by

$$f \mapsto \int k(x, y) f(y) d\mu(y) \tag{5.8}$$

This operator is adjoint to the embedding operator $R : \mathcal{H} \to L^2_\mu$ given by [10] $R\{f\}(x)$ equating to just $f(x)$, so we have $\langle f, Tg \rangle_{\mathcal{H}} = \langle Rf, g \rangle_{L^2_\mu}$.

We will assume that we have a set $\mathcal{D}$ consisting of N unlabelled data points. We wish to label a subset of $\mathcal{D}$, denoted S, of cardinality $M < N$ such that the total variation $\mathbb{E}_\mu \left[|p_{y|x} - g^*| \right]$ [1] is small for an optimal function g^* learned on S. We will use two sets of indices here. The first, $\mathcal{I} : \{1, 2, ..., M\} \to S$ indexes the selected points. The second, $\mathcal{J} : \{1, 2, ..., N\} \to \mathcal{D}$ indexes all (labelled or unlabelled) points. The notation x_{i_l} refers to $\mathcal{I}(l)$: the l^{th} data point under index $\mathcal{I}$. The notation x_{j_k} refers to $\mathcal{J}(k)$: the k^{th} data point under index $\mathcal{J}$. These indices need not coincide on the selected data points.

We will also introduce the following matrix building notation: if q is an index with domain $\{1, 2, \cdots, Q\}$ and p is an index with domain $\{1, 2, \cdots, P\}$, then

$$\left[a_{p_l} \right]^l \triangleq \begin{bmatrix} a_1 & a_2 & \cdots & a_P \end{bmatrix} \tag{5.9}$$

$$\left[a_{p_l} \right]_l \triangleq \begin{bmatrix} a_1 & a_2 & \cdots & a_P \end{bmatrix}^T \tag{5.10}$$

$$\left[[a_{p_l q_{l'}}]_l \right]^{l'} = \left[[a_{p_l q_{l'}}]^{l'} \right]_l = \left[a_{p_l q_{l'}} \right]_l^{l'} \tag{5.11}$$

where the final three are all given by the matrix whose ij^{th} element is given by $a_{p_i q_j}$, $1 \leq i \leq P,\ 1 \leq j \leq Q$.

Finally, the total variation is a 1-norm. We note that $p_{y|x}$ is an element of L^1_μ, as

$$\int \left| p_{y|x}(x) \right| d\mu(x) = p(Y = 1) \leq 1 < \infty \tag{5.12}$$

To begin estimating $\delta_{\hat{p}}$ under the selected training set, we will begin with a definition.

[1] By writing the total variation in this form, we are implicitly assuming that our problem is 2-class. While this is not always the case, it makes the theory notationaly convenient. The theory can be extended to multiple classes by using a diagonal matrix kernel such as $K(x, y) = k(x, y) I_{\mathcal{Y}}$.

Definition 10. *Let V be a subspace of $\mathcal{H}$. Then the power function on V, denoted P_V, is the function whose point-wise evaluation is given by*

$$P_V(x) = \sup_{\|f\| \leq 1} |f(x) - \mathit{proj}_V \{f\}(x)| \tag{5.13}$$

where proj_V is the orthogonal projection operator onto V.

We will deal with the particular finite dimensional subspace $V_S = \text{Span}\left(\{k(\cdot, x_{i_l})\}_{l=1}^{M}\right)$. This is the span of kernel translates. When V is such a subspace, the projection operator proj_V takes on the following well known result in approximation theory [10].

Lemma 14. *Let*

$$K_{SS} \triangleq \left[k(x_{i_l}, x_{i_{l'}})\right]_{l}^{l'} \tag{5.14}$$

and for any $x \in \mathcal{X}$, let

$$K_{xS} = K_{Sx}^T \triangleq \left[k(x, x_{i_l})\right]^{l} \tag{5.15}$$

Then for all $f \in \mathcal{H}$,

$$\mathit{proj}_{V_S} \{f\}(x) = K_{xS} K_{SS}^{-1} \left[f(x_{i_l})\right]_{l} \tag{5.16}$$

In particular, $\text{proj}_{V_S}\{f\}$ is the smallest function in V_S (w.r.t. RKHS norm $\|\cdot\|_{\mathcal{H}}$) that agrees with f at the location of the selected data points.

We will next cite another theorem from approximation theory [10]. This theorem uses the power function to bound the error induced in projecting an arbitrary function in $\mathcal{H}$ onto V_S.

Lemma 15. *Let $f \in \mathcal{H}$. Then*

$$|f(x) - \mathit{proj}_{V_S}\{f\}(x)| \leq |P_{V_S}(x)| \|f\|_{\mathcal{H}}. \tag{5.17}$$

We would like to apply lemma 15 to $p_{y|x}$. But we cannot because we didn't assume that $p_{y|x}$ is in $\mathcal{H}$. Instead, we will decompose $p_{y|x}$ into a part in $\mathcal{H}$ and a part not in $\mathcal{H}$. To do this, we will use the operator T to write:

$$p_{y|x} = (I - RT)\{p_{y|x}\} + RT\{p_{y|x}\} \tag{5.18}$$

$$\|p_{y|x}\|_{L^1_\mu} \leq \|(I - RT)\{p_{y|x}\}\|_{L^1_\mu} + \|RT\{p_{y|x}\}\|_{L^1_\mu} \tag{5.19}$$

In which we have a decomposition of error terms in (5.19). Note that $p_{y|x}$ is a valid input to T since $p_{y|x} \in L^1_\mu \subset L^2_\mu$ when $\mu(X) \leq 1$. There is not much that can be done about the first of these terms except to use an expressive RKHS. We will however assume that we can bound this term by a small error dependent on that RKHS.

$$\|(I - RT)\{p_{y|x}\}\|_{L^1_\mu} \leq \epsilon_{\mathcal{H}} \tag{5.20}$$

Next, we bound the RKHS norm of $T[p_{y|x}]$.

Lemma 16. $\|T\{p_{y|x}\}\|_{\mathcal{H}} \leq p(y = 1)$

Proof.

$$\begin{aligned}
\|T\{p_{y|x}\}\|^2_{\mathcal{H}} &= \langle T\{p_{y|x}\}, T\{p_{y|x}\}\rangle_{\mathcal{H}} \\
&= \langle RT\{p_{y|x}\}, p_{y|x}\rangle_{L^2_\mu} \\
&= \int \left(\int k(x, x')p_{y|x}(x')d\mu(x') \right) p_{y|x}(x)d\mu(x) \\
&= \int \int k(x, x')p_{y|x}(x)p_{y|x}(x') \left(d\mu(x) \otimes d\mu(x')\right) \\
&\leq \int \int p_{y|x}(x)p_{y|x}(x') \left(d\mu(x) \otimes d\mu(x')\right) \\
&= \left(\int p_{y|x}(x)d\mu(x)) \right)^2 = p^2(y = 1)
\end{aligned} \tag{5.21}$$

□

Then combining lemmas 15 and 16, we conclude:

$$\left|T\{p_{y|x}\}(x) - \text{proj}_{V_S}\{Tp_{y|x}\}(x)\right| \leq |P_{V_S}(x)|\, p(y=1) \tag{5.22}$$

Next we will perform some manipulations to calculate empirical L^1_μ norms of $P_{V_S}(x)$ in a nice form. We begin with our final cited theorem from approximation theory [56].

Lemma 17. *Let k_{V_S} be the double projection of the kernel function $k(\cdot,\cdot)$ into V_S. That is,*

$$k_{V_S}(x,y) = proj^x_{V_S} \circ proj^y_{V_S}\{k(\cdot,\cdot)\}(x,y) \tag{5.23}$$

where the rightmost projection occurs in the second argument and the leftmost projection occurs in the first argument. Then

$$|P_{V_S}(x)| = \sqrt{K(x,x) - K_{V_S}(x,x)} \tag{5.24}$$

We will need to prove one more lemma before presenting the main results. The lemma relates evaluations over doubly-projected functions to evaluations over the original unprojected function.

Lemma 18. *Let $\mathcal{Z}$ be any indexing of P elements of $\mathcal{D}$. Let $ev_{\mathcal{Z}}$ be the evaluation operator on functions of two arguments yielding*

$$f \mapsto \left[f(x_{z_l}, x_{z_{l'}})\right]^{l'}_{l} \tag{5.25}$$

and let $K_{DS} = K^T_{SD} = \left[K_{x_{j_l} S}\right]_l$, then:

$$ev_{\mathcal{J}} \circ proj^{x}_{V_S} \circ proj^{y}_{V_S}\{f\} = K_{DS}K_{SS}^{-1}ev_S\{f\}K_{SS}^{-1}K_{SD} \tag{5.26}$$

Proof. The rightmost projection in the expression $proj^{x}_{V_S} \circ proj^{y}_{V_S}\{f\}(x,y)$ occurs in the second argument. That is, it is the projection of the function $f(x,\cdot)$ which views x as fixed. Then by lemma 14, we have (swapping the order of vectors in the quadratic form):

$$proj^{y}_{V_S}\{f\}(x,y) = \Big[f(x,x_{i_l})\Big]^{l} K_{SS}^{-1}K_{Sy} \tag{5.27}$$

Now, we can re-write this as

$$proj^{y}_{V_S}\{f\}(x,y) = \sum_{l=1}^{M} c_l(y)\ f(x,x_{i_l}) \tag{5.28}$$

for some coefficients $c_l(y),\ l = 1,2,...,M$. Then by linearity of the orthogonal projection operator, we have

$$proj^{x}_{V_S} \circ proj^{y}_{V_S}\{f\}(x,y) = \sum_{l=1}^{M} c_l(y)\ proj^{x}_{V_S}\{f(x,x_{i_l})\}(x) \tag{5.29}$$

or equivalently by (5.27),

$$proj^{x}_{V_S} \circ proj^{y}_{V_S}\{f\}(x,y) = \Big[proj^{x}_{V_S}\{f(x,x_{i_l})\}(x)\Big]^{l} K_{SS}^{-1}K_{Sy} \tag{5.30}$$

Then applying lemma 14 column-wise, we obtain:

$$\begin{aligned}\Big[proj^{x}_{V_S}\{f(x,x_{i_l})\}(x)\Big]^{l} &= \Big[K_{xS}K_{SS}^{-1}\Big[f(x_{i_{l'}},x_{i_l})\Big]_{l'}\Big]^{l} \\ &= K_{xS}K_{SS}^{-1}\Big[f(x_{i_{l'}},x_{i_l})\Big]^{l}_{l'}\end{aligned} \tag{5.31}$$

But $\left[f(x_{i_{l'}}, x_{i_l})\right]_{l'}^{l} = \left[f(x_{i_l}, x_{i_{l'}})\right]_{l}^{l'}$ is just $\text{ev}_{\mathcal{I}}\{f\}$. Then combining (5.27) and (5.31), we obtain:

$$\text{proj}_{V_S}^{x} \circ \text{proj}_{V_S}^{y}\{f\}(x, y) = K_{xS}K_{SS}^{-1}\text{ev}_{\mathcal{I}}\{f\}K_{SS}^{-1}K_{Sy} \tag{5.32}$$

Finally, to evaluate this over $\mathcal{J}$, we write:

$$\text{ev}_{\mathcal{J}}[f] = \left[K_{x_{j_l}S}K_{SS}^{-1}\text{ev}_{\mathcal{I}}\{f\}K_{SS}^{-1}K_{Sx_{j_{l'}}}\right]_{l}^{l'} \tag{5.33}$$

which factorizes to

$$\text{ev}_{\mathcal{J}}[f] = \left[K_{x_{j_l}S}\right]_{l} K_{SS}^{-1}\text{ev}_{\mathcal{I}}\{f\}K_{SS}^{-1}\left[K_{Sx_{j_l}}\right]^{l} \tag{5.34}$$

□

For the final result, we will introduce one last indexing $\mathcal{U} : \{1, 2, \cdots, N - M\} \to U$ where U is the set of unlabelled data points (the $N - M$ points not chosen). Again, x_{u_l} is shorthand for $\mathcal{U}(l)$.

Lemma 19. *Let $\hat{\mu}_{\mathcal{D}}$ be the empirical measure over $\mathcal{D}$. Let*

$$K = \left[k(x_{j_l}, x_{j_{l'}})\right]_{l}^{l'} \tag{5.35}$$

Then

$$\mathbb{E}_{\hat{\mu}_{\mathcal{D}}}\left[|P_{V_s}(x)|\right] = \frac{1}{N}\textit{Trace}\left(\sqrt{K/K_{SS}}\right) \tag{5.36}$$

where the notation X/A refers to the schur complement of X with respect to A, and $\sqrt{\cdot}$ refers to taking the element-wise square root of the matrix in its argument.

Proof. From lemma 17, we have

$$\mathbb{E}_{\hat{\mu}_{\mathcal{D}}}\left[|P_{V_s}(x)|\right] = \frac{1}{N} Trace\left(\sqrt{\text{ev}_{\mathcal{J}}[k] - \text{ev}_{\mathcal{J}}[k_{V_S}]}\right) \tag{5.37}$$

Clearly, $\text{ev}_{\mathcal{J}}[k]$ is just K. Furthermore, by lemma 18 and noting that $\text{ev}_{\mathcal{I}}[k] = K_{SS}$, we have

$$\text{ev}_{\mathcal{J}}[k_{V_S}] = K_{DS}K_{SS}^{-1}K_{SD} \tag{5.38}$$

Next, consider a *selected* point x^*. Then x^* has both a $\mathcal{J}$ index and an $\mathcal{I}$ index. Without loss of generality, suppose that its index under $\mathcal{J}$ is p and its index under $\mathcal{I}$ is q. That is, $x^* = x_{j_p} = x_{i_q}$. Then K_{Sx^*} is the q^{th} column of K_{SS}. Thus $K_{SS}^{-1}K_{Sx^*}$ is just $\mathbf{e}_q$, the vector with 1 at the q^{th} element and zeros elsewhere. Thus we have the following sequence of equalities:

$$k_V(x^*, x^*) = K_{x^*S}K_{SS}^{-1}K_{Sx^*} = k(x^*, x_{i_q}) = k(x^*, x^*) \tag{5.39}$$

We can thus ignore the terms in (5.24) corresponding to selected points. Now Let

$$K_{UU} = \left[k(x_{u_l}, x_{u_{l'}})\right]_l^{l'} \tag{5.40}$$

$$K_{US} = K_{SU}^T = \left[k(x_{u_l}, x_{i_{l'}})\right]_l^{l'} \tag{5.41}$$

Then the diagonal of K_{UU} contains all terms of the form $k(x, x)$ where x is not a selected point, and the diagonal of $K_{US}K_{SS}^{-1}K_{SU}$ contains all terms of the form $k_V(x, x)$ where x is

not a selected point. Thus we have

$$\mathbb{E}_{\hat{\mu}_{\mathcal{D}}}\left[|P_{V_s}(x)|\right] = \frac{1}{N} Trace\left(\sqrt{K_{UU} - K_{US}K_{SS}^{-1}K_{SU}}\right) \tag{5.42}$$

whose argument can be recognized as the Schur complement referred to in (5.36), completing the proof. □

Finally, we will combine all of these lemmas with the following assumption:

Assumptions 2. *g^* performs better than $proj_{V_S}p(y|x)$ in terms of conditional total variation.*

From which we conclude the following theorem.

Theorem 12. *Let $\delta_{\hat{p}}^{emp}$ be the empirical (monte-carlo) estimate of $\delta_{\hat{p}}$. Then under the assumptions of this subsection:*

$$\delta_{\hat{p}}^{emp} \leq \frac{p(y=1)}{N} Trace\left(\sqrt{K/K_{SS}}\right) + \epsilon_{\mathcal{H}} \tag{5.43}$$

For notational purposes, we will expand the trace term explicitly to:

$$\frac{p(y=1)}{N} \sum_{l=1}^{N-M} \sqrt{1 - K_{Sx_{u_l}}^T K_{SS}^{-1} K_{Sx_{u_l}}} \tag{5.44}$$

Analytical Investigation of $\epsilon_{\mathcal{H}}$ and Choosing a Good Kernel

$\epsilon_{\mathcal{H}}$ can be explicitly calculated as follows: letting (ϕ_i, λ_i) denote the i^{th} eigenvector and eigenvalue of T, we have:

$$\|(I - RT)\{p_{y|x}\}\|_{L_\mu^1} = \| \sum_i (1 - \lambda_i) \left\langle p_{y|x}, \phi_i \right\rangle_{L_\mu^2} \phi_i \|_{L_\mu^1} \tag{5.45}$$

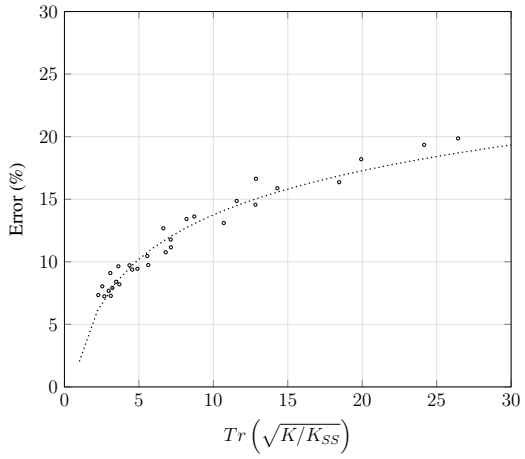

Figure 5.1: Caption

where the argument inside the L^1_μ norm can be viewed as a high pass filtering of $p_{y|x}$ in the kernel space. Picking a good kernel for use with this metric amounts to two conditions. First, we would like the above term $\epsilon_{\mathcal{H}}$ to be small. This can be obtained by taking a small sample of training data and evaluating a monte carlo estimate of the above term, using a discretized version of the kernel eigenvalue problem.

5.3.1 Converting to multiple classes

Converting this bound to multiple classes is as simple as removing the $p(y = 1)$ term and dividing by 2. This is because the total variation over multiple classes is given by half the sum of each L_1-norm. Thus the $p(y = c)$ terms in the bound of each L_1-norm sum together to 1, and we are just left with the remaining $\frac{1}{2}$. $\epsilon_{\mathcal{H}}$ can be similarly estimated by performing the estimation via equation (5.45) for each class variable, and then summing and dividing by 2.

5.4 Experiments

5.4.1 Correspondence Between Bound and Classification Accuracy

We test if our bound has any generalizable meaning to classification accuracies. To do so, we first took the MNIST with the cosine kernel and swept over training data sizes from 5% to 90% and plotted the trace term against the classification accuracy of a fully connected feed forward neural network with 1000 hidden units in Figure 5.1 (right). We see a strong correspondence between these two variables. However, to ensure that this correspondence is meaningful, we also need to ensure that it exists when the training data size is controlled - since both terms, in isolation, decrease as M grows.

To do this, we took several samples of training data over a variety of datasets provided by OpenML [100] and trained them on a fully connected feed-forward neural network with 1000 hidden units - plotting the trace term against the classification error in each case. In each case, we took 20% of the full dataset as training data. We have provided the resulting scatter plots in Figure 5.2 which show that, in this controlled scenario, the two variables are still correlated. Each point in these plots corresponds to a different ratio of selected points to random points in that fixed-size training dataset, where selection is done with an 'inverse' heuristic method which just picks data corresponding to small diagonal elements in K^{-1}.

A second way of ensuring that the relationship found in Figure 5.1 is meaningful is to observe the behavior of different data selection methods over a training data sweep. We have done this in Figure 5.3. On the left hand side of Figure 5.3, we have plotted training data size against the trace term (for MNIST under the cosine kernel) for five such methods of training data selection: random selection, facility location, uncertainty sampling, and the inverse heuristic from the last paragraph. On the right hand side of Figure 5.3, we have a similar plot where the trace term is replaced by classification error of a fully connected feed forward neural network with 1000 hidden units. We see that the behavior of the trace term

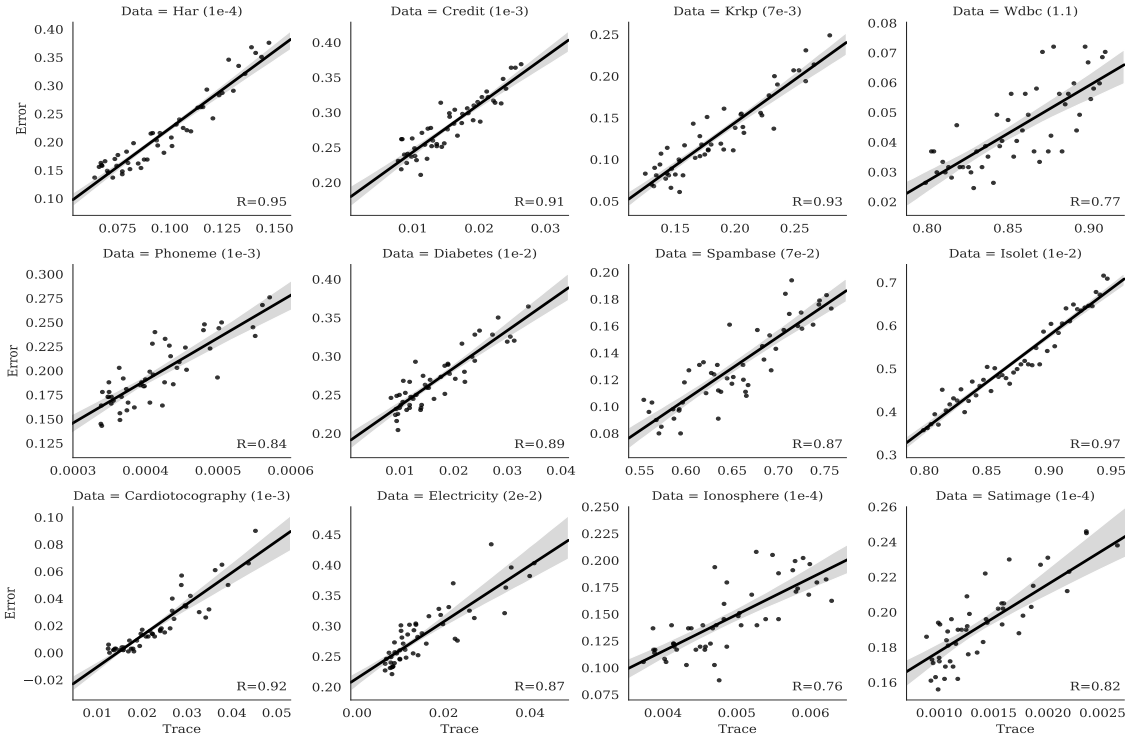

Figure 5.2: Classification errors against the data quality measure with a fixed training data size for varying datasets. Dataset and corresponding rbf γ value are indicated at the top of each plot.

plots are carried over to the classification error plots. Some small scale information is lost, mostly due to the fact that the error plot is more noisy, but the main global properties are intact. This correspondence of behavior further shows that there is a link, independent of training data size, between our bound and classification error.

5.5 Chapter Summary

This chapter has provided a novel information theoretic perspective on active learning methods. It has provided an information theoretic proof of the viability of the facility location function data selection method, and derived a new information theoretic bound which is highly applicable to evaluating and analyzing other active learning strategies. Experiments show that this bound is quite tight, and that it is indicative of dataset quality in

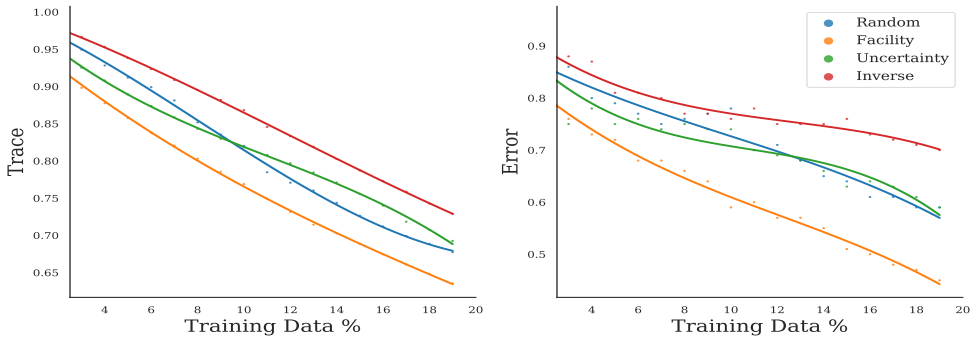

Figure 5.3: MNIST data quality measure and classification error against training data size for several methods of training data selection.

terms of classification accuracies.

CHAPTER 6

MITIGATING INFORMATION LOSSES IN PRACTICE

6.1 Phase identification Properties and Similar Problems

Primary to our applications of the above theoretical framework are the following three properties pertaining to the dataset/classifier pair $(\mathcal{X}, \mathcal{F})$.

- $(\mathcal{X}, \mathcal{F})$ is inherently low entropy. That is, the Maximum Mutual Information of $(\mathcal{X}, \mathcal{F})$, which we will define soon, is relatively small.
- $\mathcal{X}$ is small to medium in size.
- Obtaining labels for $\mathcal{X}$ requires slow, costly field projects.

The first property is required for the information-loading method which we will describe soon. The second is necessary for the inverse-schur heuristic of data selection from the previous chapter. The third is not necessary, but encourages the use of active learning methods which are central to our solution strategies.

These properties are not limited to Phase Identification. Indeed, the first is likely to be true whenever a problem inherits a well-fit linearized model common to several cyber physical systems. The second is true when a problem domain is limited to small cities or subsets of large cities. The third is often true when data requires physical measurements to obtain.

Thus, while we have limited our experiments to the phase identification problem here, the implications are much larger, spanning several problems in cyber physical systems.

6.2 Technique - Inverse Schur Training Data Selection

6.2.1 Motivation - Field Testing

The first technique addresses the technical limitations of supervised phase identification. Namely, that obtaining labeled data is time consuming. Obtaining phase labels for a given customer requires on-site measurements with phasor measurement unit or phasing meter. Gathering phase connection information for a large number of customers with these equipment to serve as labels would be prohibitively costly.

It is critical that we get as much as we can out of just a few phase labels. This inspires the use of active learning [86, 98]. This is a field of machine learning which focuses on the selection of training data points that best represent the whole data set. It can be loosely divided into unsupervised methods and supervised methods - with the latter division dominating the number of available techniques by a large margin. Supervised methods generally update a machine learning model in conjunction with label selection. But this step is lengthy on its own, and can not be performed in parallel with on-location travelling or with device installation. Thus supervised techniques may be deemed too lengthy when data acquisition requires field testing. Thus we will focus on unsupervised techniques. We will build upon some recent theoretical work on the link between information losses and data selection [32].

6.2.2 Training Data Selection via Information Losses

From the previous chapter, we see that we can reduce information losses by minimizing the term $Trace(\sqrt{K/K_{SS}})$ under a well fitting kernel (e.g. such that the term $\epsilon_{\mathcal{H}}$ is small). For the phase identification problem, and for problems whose datasets are generated by models with well-fitting linearizations, the cosine kernel works well. Thus we wish to solve the following optimization problem:

$$\min_S Trace\sqrt{K/K_{SS}} \tag{6.1}$$

$$s.t.\ k(x,y) = \frac{x^T y}{\|x\|\|y\|} \tag{6.2}$$

However, strict optimization of (6.1) will take exponential time. Furthermore, since each evaluation of K/K_{SS} for a given S will require a matrix inverse calculation, greedy optimization performs poorly as well. Indeed, the computational complexity grows with $O(NM^4)$. If we choose M such that $M = \rho N$ for some proportion ρ, then this scales with $O(N^5)$, which is quite poor.

However, we can heuristically optimize (6.1) in $O(N^3)$ time as follows: First, we note that

$$K^{-1} = \begin{bmatrix} \cdot & \cdot \\ \cdot & (K/K_{SS})^{-1} \end{bmatrix} \tag{6.3}$$

Thus we can control the value of $Trace((K/K_{SS})^{-1})$ by data point selection in a predictable way. That is, the trace of $(K/K_{SS})^{-1}$ will be large if we pick points corresponding to small diagonal elements in K^{-1}.

We then note a few correspondences between the trace of a matrix and the trace of its inverse. First, by the Cauchy-Schwartz inequality, we have:

$$len(A) \leq Tr(A)Tr(A^{-1}) \tag{6.4}$$

where $len(A)$ is the size of either axis of A.

Furthermore, we have the following lemma:

Lemma 20. *Consider the partially ordered set of matrices $M = \{A : A = K/\{\alpha\}$ for some index α of size M in the Loewner order. That is, $A \leq B$ iff. $A - B$ is positive semi-*

definite. Let A^* be the matrix in this poset corresponding to the index α^* which maximizes $(K/\{\alpha^*\})^{-1}$. Then there exists no matrix $B \in M$ such that $A^* > B$.

This lemma, which we will prove shortly, shows that picking data points to maximize $Tr((K/K_{SS})^{-1})$ will force $Tr(K/K_{SS})$ to be smaller than that of a large range of other datasets. Specifically, if B is comparable to A^* in the Loewner order (i.e. either $A^* - B$ or $B - A^*$ is positive semidefinite) then A^* must be the smaller of the two and so $Tr(A^*) \leq Tr(B)$ by monotonicity of the trace operator. We now proceed to this lemma's proof.

Proof. By the hypothesis of the lemma, A^* has the largest inverse trace of all matrices in M. That is, $Tr(C^{-1}) \leq Tr((A^*)^{-1})$ for all $C \in M$. Now suppose for the sake of contradiction that there exists $B \in M$ such that $A^* > B$. Then $(A^*)^{-1} < B^{-1}$ by monotanacity of the inverse operator. This, in turn, implies that $Trace((A^*)^{-1}) < Trace(B^{-1})$, contradicting the hypothesis. □

6.3 Technique - Information Loading

6.3.1 Voltage Data and Phase Identification

The second technique exploits the properties of our feature space. We will first describe what our feature space is and why we use it. We will then move into deriving the exploitable properties of this space - the main assertion being that a standardized voltage dataset has relatively low entropy. We will then experimentally validate this assertion. Finally, we will describe a technique, called information loading, which will exploit this property.

Our feature space consists of smart-meter voltage magnitude time series data. Voltage data is fairly informative of phase type. Thus it is a good candidate feature space for the phase identification problem. This can be seen through the following example.

Consider a power injection at bus k whose phase type is AB. This induces a current along the lines A and B. Thus, a voltage change will occur along those lines throughout

the circuit. Any customer also feeding from either of those lines will notice a change. Due to the capacitive and inductive effects of the primary feeder, both lines will also induce a voltage change along the lines C and n. However, the off-diagonal elements of the phase impedance and shunt admittance matrices are much smaller than the diagonal ones. Hence, the power injection at bus k will have much less effect on phase C than phase A and B. Thus, the customers whose voltage data is most effected by this power injection are those on AB. Second to these customers are customers who share either the phase line A or B (An, Bn, BC, CA). Finally, customers who do not pull from either A or B (Cn) will hardly be effected at all.

Thus, while this current injection somewhat affects all customers, it affects customers of the same phase the most. Hence, voltage data is informative of phase connection type.

6.3.2 Properties of Voltage Data

We will now provide some rough quantitative descriptions of distribution systems. We consider a simple model in which each 'customer' corresponds to a distribution transformer branched from the primary feeder. That is, we are ignoring the properties of the secondary distribution. We assume that bus 0 represents the distribution substation and that each bus is a fixed electrical distance Δ_E from the previous bus. We will further assume that the impedance along the primary feeder is a constant z. We assume without loss of generality that the customers are indexed in order by their electrical distance from the distribution transformer. We let $V_p(x)$ denote the voltage on phase p at position x. We assume that we are working in a per unit system and that our system is fairly balanced, such that $V_A(0) \approx 1$, $V_B(0) \approx e^{-\frac{2\pi}{3}}$, $V_C(0) \approx e^{\frac{2\pi}{3}}$, $V_n(0) \approx 0$. We will assume for simplicity that there are no three phase loads in the circuit.

Now, let μ denote the degenerate measure $\sum_{j=1}^{N} \delta(j\Delta_E)$ and let i denote a stochastic process over $[0, N\Delta_E]$ taking complex values. When i is evaluated at any $j\Delta_E, j =$

$1, 2, \cdots, N$, this is to be interpreted as the current injections into the system by customer j. Finally, we denote as $\vec{\phi}$ the function defining the phase of the current injection i. That is, if the current injection at position x' is on phase An, then $\phi(x') = \begin{bmatrix} 1 & 0 & 0 & -1 \end{bmatrix}^T$ whereas for phase AB we would have $\phi(x') = \begin{bmatrix} 1 & -1 & 0 & 0 \end{bmatrix}^T$.

Now, denoting $\vec{V}(x) \triangleq \begin{bmatrix} V_A(x) & V_B(x) & V_C(x) & V_n(x) \end{bmatrix}^T$, we have:

$$\vec{V}(x) \approx \vec{V}(0) - z \int_0^x x' i(x') \vec{\phi}(x') d\mu(x') \tag{6.5}$$

Now, the voltage measured at x by a meter, $\tilde{V}(x)$, is given by $\langle \vec{\phi}(x), \vec{V}(x) \rangle$. Thus:

$$\begin{aligned} &\tilde{V}(x = k\Delta_E) \\ &\approx \langle \vec{\phi}(x), \vec{V}(0) \rangle - z \int_0^x x' i(x') \, \langle \vec{\phi}(x), \vec{\phi}(x') \rangle \, d\mu(x') \\ &= \langle \vec{\phi}(x), \vec{V}(0) \rangle - z\Delta_E \sum_{j \le k} j \cdot i(j\Delta_E) \, \langle \vec{\phi}(x), \vec{\phi}(j\Delta_E) \rangle \end{aligned} \tag{6.6}$$

In which we already see our intuition pop out: the customers contributing most to the voltage measurement at x are the customers whose $\vec{\phi}$ vectors have the largest inner product with $\vec{\phi}(x)$, i.e., the customers who share phase lines with the customer at x.

Now, (6.6) refers to phasor quantities. However, smart meters typically only return time averaged voltage magnitudes. We can take this into account by modifying each $\vec{\phi}$ vector in correspondence with the assumption that the circuit is fairly balanced. As such, we imagine lumping all of the single phase customers into one large wye-connected load and all of the two phase customers into one large delta-connected load. We consider each customer's current injections as contributing to current injections on these loads. Then the effect of all of the A injections, for example, is to drop the voltage magnitude V_A along the primary feeder, but not effect the neutral line at all. Thus, the $\vec{\phi}$ corresponding to phase

A should be modified to $\tilde{\phi} = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$. Similarly, the $\vec{\phi}$ corresponding to phase AB should be modified to $\tilde{\phi} = \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & 0 & 0 \end{bmatrix}$ to account for the transformation from a line-line current magnitude into a line-neutral current magnitude. With all of the above approximations, and denoting as $\bar{V}$ the time-averaged version of $\tilde{V}$ we have

$$\bar{V}(k\Delta_E) \approx g(k\Delta_E) - |z|\Delta_E \sum_{j \leq k} j \cdot |i(j\Delta_E)| c_{k,j} \tag{6.7}$$

where $g(k\Delta_E)$ is 1 if customer k is single phase and $\sqrt{3}$ if customer k is two phase. $c_{k,j}$ is an inner product with $\frac{1}{\sqrt{3}} \leq |c_{k,j}| \leq \frac{2}{\sqrt{3}}$. We can tighten the lower bound on the diagonal such coefficients as $1 \leq c_{i,i}$.

Collecting all $\bar{V}(k\Delta_E)$ together into one vector $\bar{\mathbf{V}}$, we have:

$$\bar{\mathbf{V}} = \mathbf{g} - |z|\Delta_E \mathbf{CJI} \tag{6.8}$$

where $\mathbf{g}$ collects the g functions over the customers, $\mathbf{I}$ collects current magnitudes over each customer, $\mathbf{J} = diag(1, 2, \cdots, N)$, and

$$\mathbf{C} = \begin{bmatrix} c_{11} & 0 & 0 & \cdots & 0 \\ c_{21} & c_{22} & 0 & \cdots & 0 \\ c_{31} & c_{32} & c_{33} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{N1} & c_{N2} & c_{N3} & \cdots & c_{NN} \end{bmatrix} \tag{6.9}$$

Thus the statistics of $\bar{\mathbf{V}}$ are inherited linearly from those of $\mathbf{I}$. Assuming that I is an iid Gaussian distribution with covariance σ^2, this yields a multivariate Gaussian distribution for $\bar{\mathbf{V}}$. The covariance of $\bar{\mathbf{V}}$ is given by $\Sigma = |z|^2 \Delta_E^2 \sigma^2 (\mathbf{CJ})(\mathbf{CJ})^T$. We note that this iid

Gaussian assumption is not prohibitive when our goal is in showing that our entropy is low; iid Gaussian random variables *maximize* entropy at a fixed variance level. Thus, if the iid Gaussian assumption of $\mathbf{I}$ does not hold, then the entropy of the dataset will be even lower than what we predict.

6.3.3 Entropy Analysis

The above approximate derivation allows us to analyze the entropic properties of smart-meter voltage measurements.

Lemma 21. *For a distribution system following the model of subsection 6.3.2, we have:*

$$ln\left(\frac{e}{N}\right) - \frac{1}{N}ln\left(\frac{\sqrt{\frac{2\pi}{e^2}}}{(N+1)}\right) \leq \frac{H(\bar{\mathbf{V}})}{N}$$
$$\frac{H(\bar{\mathbf{V}})}{N} \leq \frac{1}{2}ln\left(\frac{12e}{N}\right) - \frac{1}{N}ln\left(\frac{\sqrt{\frac{e^2}{4\pi}}}{(N+1)}\right) \tag{6.10}$$

Proof. We first note the growth rate of the diagonal elements of $\bar{\mathbf{V}}$'s covariance matrix Σ.

$$\Sigma_{11} \propto c_{11}^2 \tag{6.11}$$
$$\Sigma_{22} \propto c_{21}^2 + 4c_{22}^2 \tag{6.12}$$
$$\Sigma_{33} \propto c_{31}^2 + 4c_{32}^2 + 9c_{33}^2 \tag{6.13}$$

And thus we can estimate:

$$\frac{1}{3}\frac{k(k+1)(2k+1)}{6} \leq \frac{\Sigma_{kk}}{\alpha} \leq \frac{4}{3}\frac{k(k+1)(2k+1)}{6} \tag{6.14}$$

where $\alpha = |z|^2\Delta_E^2\sigma^2$. This estimate is important due to the preprocessing of our dataset. When applying a machine learning algorithm, we typically scale every data-set down such

that the diagonal elements of the covariance matrix are equal to 1. This makes the effective covariance $\tilde{\Sigma} = D\Sigma D$ for some matrix D such that the diagonal of $\tilde{\Sigma}$ consists of $1's$.

We can now estimate the entropy of a scaled voltage dataset via the entropy of a Gaussian random variable.

$$H(X) = \frac{1}{2} ln \left((2\pi e)^N det(\tilde{\Sigma}) \right) \tag{6.15}$$

$$= \frac{1}{2} ln \left((2\pi e)^N det(\alpha DCJ^2C^T D) \right) \tag{6.16}$$

Now, for calculating the determinant portion, we note that

$$det(\alpha DCJ^2C^T D) = \alpha^N det(D)^2 det(C)^2 det(J)^2 \tag{6.17}$$

First, we have $det(J)^2 = N!^2$. Further, we can bound $det(C)$ as:

$$1 \leq det(C)^2 \leq \left(\frac{4}{3} \right)^N \tag{6.18}$$

where the lower bound occurs if every customer is two-phase attached and the upper bound occurs if every customer is single-phase attached. Finally, due to (6.14) and the scaling property:

$$\frac{9^N}{N!(2N+1)!(N+1)} \leq det(\alpha D)^2 \leq \frac{36^N}{N!(2N+1)!(N+1)} \tag{6.19}$$

Thus

$$\frac{9^N N!}{\alpha^N (3N)!(N+1)} \leq det(D)^2 det(J)^2 \leq \frac{36^N N!}{\alpha^N (2N)!(N+1)} \tag{6.20}$$

Now, by Sterling's approximation, we can further obtain:

$$\frac{\sqrt{\frac{2\pi}{e^2}}\left(\frac{e^2}{\alpha N^2}\right)^N}{(N+1)} \leq det(D)^2 det(J)^2 \leq \frac{\sqrt{\frac{e^2}{4\pi}}\left(\frac{9e}{\alpha N}\right)^N}{(N+1)} \tag{6.21}$$

From which we can finally obtain:

$$\begin{aligned} Nln\left(\frac{e}{N}\right) - ln\left(\frac{\sqrt{\frac{2\pi}{e^2}}}{(N+1)}\right) &\leq H(\bar{\mathbf{V}}) \\ H(\bar{\mathbf{V}}) \leq \frac{N}{2} ln\left(\frac{12e}{N}\right) - ln\left(\frac{\sqrt{\frac{e^2}{4\pi}}}{(N+1)}\right) & \end{aligned} \tag{6.22}$$

□

It is easy to see that both the lower bound and the upper bound become quite negative with N. For $N \approx 5000$, these bounds estimate the per-customer entropy to be roughly between -4 and -11 bits. If the smart meters are encoded with 16 bits, then the per-customer entropy that the machine learning algorithm 'sees' is between 5 and 12 bits. Thus we say that this problem is 'inherently low entropy'. We note that the main reason this low entropy occurs is due to the fact that the covariance matrix diagonal values scaled with the distance of the corresponding customer from the substation. Had this not been the case - i.e., had every customer been of equal uncertainty, then the entropy of the dataset would have been much larger - scaling positively as N grows large.

We obtain a low entropy dataset because the uncertainty of the customer's voltage data is dominated by those customers far from the substation. This will stay true of many problems in networked systems. Since this is the main motivation behind the second technique introduced in this chapter (information loading), we conjecture that the technique is highly applicable to problems beyond phase identification.

6.3.4 Maximum Mutual Information (MMI) Estimation: Further Evidence of the Low Entropy Feature Space Hypothesis

For the next component of this chapter, we will need a way of estimating the amount of mutual information that a given neural network *can* carry about an input variable X. This estimation is the subject of this subsection.

We have from information theory that for any random variable U,

$$H(X) = I(X;U) + H(X|U) \tag{6.23}$$

We assume that $H(X) > 0$ and suppose we have a class of conditional distributions $\mathcal{Q}$ from which to search for a variable U meant to approximate X (e.g. the hidden layer in our neural networks). Then if $\mathcal{Q}_{U|X}$ is large enough, it will contain a subset $\mathcal{Q}'$ such that $H(X|U') > 0$ for all U' whose joint distribution follows $p(x)q'(u|x)$ for some $q'(u|x) \in \mathcal{Q}'$. Then for such U', $H(X) \geq I(X;U)$, and we have equality when $H(X|U) = 0$. This motivates the estimator

$$H(X) \geq \sup_{q(u|x)\in\mathcal{Q}} I(X;U) \tag{6.24}$$

whose optimum will not occur in $\mathcal{Q}'^c$. Thus we can then adapt the MINE-f mutual information estimator [9] to obtain

$$H(X) \geq \sup_{q(u|x)\in\mathcal{Q}_{U|X}} \sup_{t\in\mathcal{F}} \int t d\mathbb{P}_{XU} - \int e^{t-1} d\left(\mathbb{P}_X \otimes \mathbb{P}_U\right) \tag{6.25}$$

where $\mathcal{F}$ is another space of functions. From here, we can simply define $\mathcal{Q}$ and $\mathcal{F}$ as spaces of functions parameterized by deep neural networks with fixed hyper-parameters. Since (6.25) will only have equality for very large $\mathcal{Q}_{U|X}$, we will give the resulting right hand side a separate name - the *maximum mutual information* (MMI) of $\mathcal{Q}_{U|X}$. This is a number which

depends on the input feature space $\mathcal{X}$ and on the space $\mathcal{Q}_{U|X}$, and is the desired amount of mutual information that a given network (with fixed hyper-parameters) can carry about the input variable.

We have experimentally conducted these MMI estimations on five real circuits with $\mathcal{F}$ defined by a feed-forward neural network with a single hidden layer of 1000 units. We have observed a maximum MMI of just 8.67 bits, and an average of just 6.21 bits. For reference, these MMIs are much lower than that of the MNIST dataset, which has an MMI of about 21 bits [79].

6.3.5 Information Loading

We will now consider the other important term contributing to $I_{loss}(S)$: $I(X;Z)$. At first glance, it would appear that *reducing* $I(X;Z)$ would be a pertinent goal. Indeed this is the premise behind the information bottleneck method [96]. However, there is a hidden trade-off here. This is because reducing $I(X;Z)$ leads to its own form of information loss through the strong data processing inequality:

$$I(Y;Z) \leq \eta I(X;Z),\ \eta \leq 1 \tag{6.26}$$

Thus we must balance the loss reduction from $I(X;Z)$ against these additional losses. Furthermore, since we are already reducing $\delta(\hat{p})$ by using the data selection frameworks above, the marginal loss reduction that can be achieved from reducing $I(X,Z)$ is reduces as well. We thus conjecture that, unless $I(X;Z)$ is large, the losses from the data processing inequality will win out. But voltage data in general has relatively low entropy as we will show. Thus $I(X;Z)$ will be low as well for all random variables Z. Thus, we should attempt to keep $I(X;Z)$ as large as possible.

However, there are losses in $I(X : Z)$ that occur naturally. First, if any stochasticity is introduced to the neural network in use, then we can view the neural network as a lossy channel. Information losses in $I(X; Z)$ will occur as a result. This can be alleviated by just not using a stochastic network, but there is a second form of $I(X; Z)$ losses that are more critical - finite data information losses. These losses come from the fact that, even though X is instantiated for every customer, our neural classifier only sees x instantiations that are accompanied by y labels. This artificially limits the amount of X data that is seen, yielding the losses in $I(X; Z)$.

But this need not be the case. If we can write an estimator of $I(X; Z)$ as a function of neural network parameters, then we can simply add in an information 'anti-regularization' term to whatever loss function we are using. That is, if $\mathcal{L}$ is our current supervised loss function (say cross-entropy), then we can modify this to

$$\mathcal{L} - \beta I(X; Z),\ \beta > 0 \tag{6.27}$$

effectively performing the *opposite* of the information bottleneck method.

We have such an estimator introduced in this chapter already: MINE-f from subsection 6.3.4. Plugging this estimator into (6.27) will yield the desired result. See figure 6.1.

6.4 Experiments

6.4.1 Data

This analysis will be performed over 5 circuits of varying complexity from Southern California Edison, Pacific Gas and Electric Company, and FortisBC. The details of these circuits are contained in the Table 6.1. Where Degree of Balance is measured by the average current coming back on the neutral line in the distribution circuit. The obtained values are

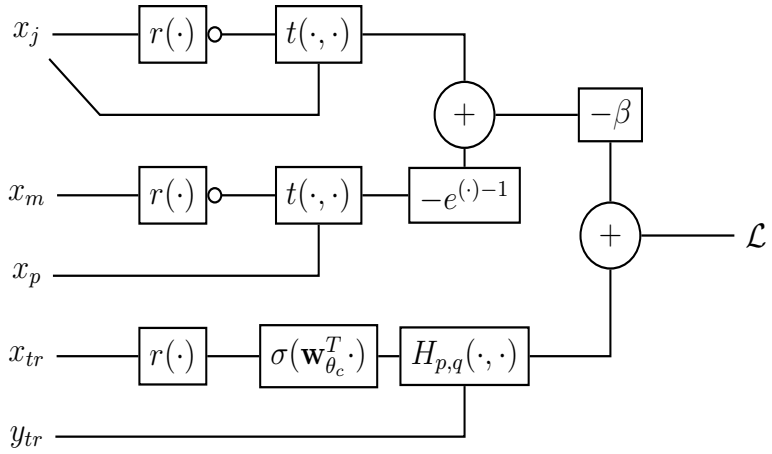

Figure 6.1: String diagram representation of the information loading forward pass. $r : \mathcal{X} \to \mathcal{Z}$ is the representation function. $t : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ is the argument of the information estimator.

Table 6.1: Distribution Circuits Characteristics

Name	$N_{consumers}$	Phase Connections	Degree of Balance
I	1892	A, B, C, ABC	Low
II	3166	All	Low
III	4629	A, B, C, ABC	High
IV	3638	B, C, AB, BC, CA	High
V	1773	A, B, C	Low

partitioned into two equal-probability bins which we denote as 'Low' and 'High'.

Each circuit contains 31 days of voltage magnitude data, sampled hourly for a feature vector of dimension 744. All experiments are performed with 5% of the total customers used as training data. Every reported number is an average over ten trials.

Empirically, circuits with more potential phase connections (e.g. A, B, C, AB, BC, CA vs. just A, B and C) typically have lower Phase Identification accuracy. This is firstly due to the fact that the difficulty of a classification task is related to the number of classes, but also due to the fact that there are nontrivial dependencies between some of these classes; for example, transformers of the AB class take current from the A line and send it back along the B line, which complicates the dynamics of transformers attached to just A or B. Balanced circuits also have lower Phase Identification accuracy than unbalanced ones, but

the effect is less significant. The more phase connections available and the more balanced the circuit is, the more 'difficult' that circuit is to identify.

6.4.2 Preprocessing

In many distribution systems, center tapped transformers are abundant. As such, voltage magnitude data will come in two bulk clusters, one near $120V$ and one near $240V$. This distinction has little relevance for the phase connection type of the corresponding customer, and will add instability into any supervised learning algorithm. We take care of this by *self*-normalizing each voltage time series such that its time-average is equal to 1.0. We follow this step with the standard preprocessing technique of *batch*-normalizing the data to have an batch-mean of 0.0 and a batch-standard deviation of 1.0.

6.4.3 Results

We first desired to establish some baseline accuracies for the phase identification problem using standard supervised learning approaches. The results of this analysis are shown in Table 6.2. We see that a two layer neural network with 500 hidden units outperforms the other methods in 4 out of 5 cases. Only on circuit II is the neural network beaten, and barely so. Thus we decided to implement our changes on this classifier.

We tested our proposed techniques in all permutations. These are all shown in Table 6.3. The first column of this table repeats the accuracy of the baseline two layer neural network from figure 6.2. The second column considers information loading in isolation. This yields minor to substantial changes. The effect is highly dependent on the circuit. In general, this technique seems to yield larger improvements to harder circuits. We then tested, training data selection under the facility location method [85]. This yields substantial improvement in every case. We observe in the fourth column that the inverse-matrix heuristic outperformed the facility location method in every case. Finally, we performed both the information

loading technique and the inverse-schur heuristic to obtain the rightmost column in the table. These combined techniques yield the highest phase identification accuracies in every case. They are phenomenally improved from the baseline results.

Circuit	Neighbors	Decision Tree	Random Forest	Neural (2-layer)
I	74.6%	71.5%	68.2%	**80.7%**
II	**64.8%**	59.3%	39.3%	64.7%
III	70.6%	59.2%	59.4%	**71.4%**
IV	67.2%	59.3%	51.8%	**75.0%**
V	41.0%	50.00%	37.00%	**51.7%**

Table 6.2: Baseline Establishment

Circuit	Baseline	I-loading	Facility	Inverse	Inverse+I-loading
I	80.7%	81.5%	86.7%	89.9%	**91.0%**
II	64.7%	80.6%	90.5%	95.4%	**96.3%**
III	74.1%	75.2%	90.6%	91.5%	**93.1%**
IV	75.0%	78.0%	91.2%	94.6%	**98.8%**
V	51.7%	59.1%	94.2%	96.1%	**97.3%**

Table 6.3: Proposed Techniques

Circuit	Correlation	Clustering	Proposed
I	37.8%	75.1%	**91.0%**
II	34.1%	56.4%	**96.3%**
III	46.4 %	65.7%	**93.1%**
IV	40.1%	53.6%	**98.8%**
V	38.4%	38.4%	**97.3%**

Table 6.4: Accuracy comparisons between the literature and the proposed method.

We've also compared the accuracies obtained by our proposed method to two of the methods of phase identification in literature. Both of these methods lie on the physical-intuition side of techniques, in contrast to the more abstract off-the-shelf machine learning

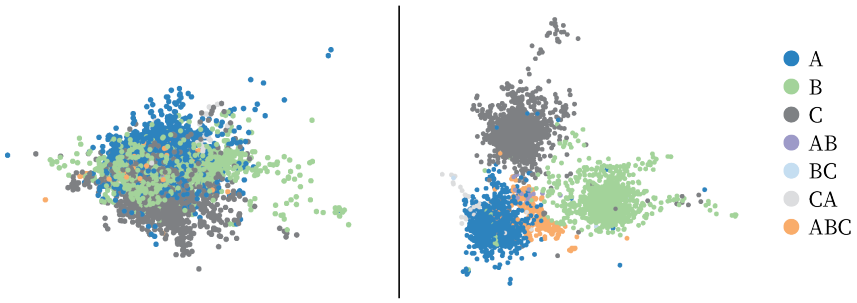

Figure 6.2: Learned representations on circuit II. (Left) random selection with no information loading. (Right) Targeted selection with information loading.

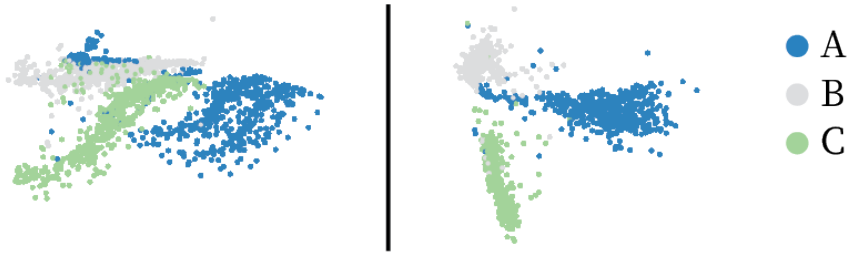

Figure 6.3: Learned representations on circuit V. (Left) random selection with no information loading. (Right) Targeted selection with information loading.

techniques. The first, which we've denoted 'correlation', slightly modifies the correlation based methods of references [62, 74, 75, 84, 107] by computing the empirical voltage correlation matrix over the customers, and using this matrix to link the customers together under complete-linkage. The second, which we've denoted 'clustering' is equivalent to that of reference [102]. We see that the literature is capable of achieving similar accuracies to the more abstract machine learning algorithms in Table 6.2, but without requiring training data. However, our proposed method, which synergistically combines both physical intuition and

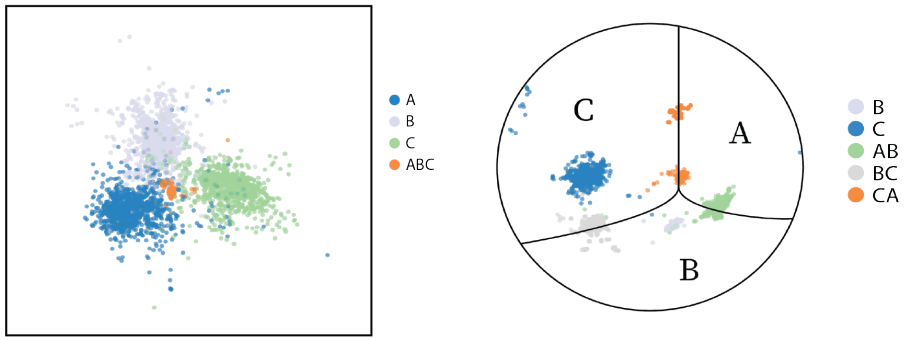

Figure 6.4: Learned representations on circuits I (left) and IV (right) with both techniques implemented.

the more abstract learning algorithms, beats both types of methods individually, yielding the best results on these datasets by a large margin.

Since the central idea of this dissertation lies in learning better representations, we ought to visualize what the representations have learned. This has been done for two of the circuits, and is visualized in figures 6.2 and 6.3. Each point on these plots corresponds to an averaged sample from the distribution $p(z_i|x_i)$ averaged over 50 trials. Each of these averages have dimension 500. We reduced this to 2 dimensions by performing PCA and projecting each point onto the first two principal components.

We observe that the representation has indeed learned much better representations by using these techniques. In the baseline case, there is little separation between the classes - especially in the case of circuit II. By implementing the proposed techniques, we see more class separation in these variables. Furthermore, we observe some learning of the relationships between the classes. For example, in circuit II, the learned representation has placed ABC directly in the middle of the classes A B and C. This is significant since, by subsection 6.3.2, phase ABC truly is a 'combination' of phases A, B, and C when it comes to voltage data. In the case of circuit V, we have learned even more. In the final representation, we see three branches of representation data - one for each class. The

distance from the central node of these representations corresponds to distance along the primary feeder of the distribution circuit.

Finally, we have plotted the final representations under both techniques for circuits I, and IV in figure 6.4. The behavior of circuit III is nearly identical to circuit I, so we are not presenting its plot for the sake of space. In the cases of circuits I and III, we see similar behavior. Independent An Bn and Cn clusters have arisen, and an ABC cluster has appeared in the middle. This is, again, what we expect from circuits consisting of those phase types. Circuit IV is a bit more interesting because it consists of phase types Bn, Cn, AB, BC and CA, with Bn having very little representation. Clusters of each phase type have appeared with decent separation, so classification will at least be easy. More interestingly, however, is the fact that the AB cluster has appeared opposite to the Cn cluster, CA opposite Bn. Assuming the location of the non-existent An cluster to the top right, these positions make much sense.

6.5 Conclusion

This chapter has used the theory of information losses to propose the application of two novel techniques - inverse schur data selection and information loading - to the phase identification problem. These techniques have synergistically combined the the abstract, problem agnostic, supervised learning techniques found in machine learning research with the physical intuitions that are often employed in more specific power systems projects. As such, we observe substantial improvements in phase identification accuracy over both the purely abstract methods and those methods which only base themselves on the physical intuitions. Furthermore, we have observed that the representations learned upon using these techniques are much more meaningful than the baseline representations, giving us highly interpret-able results which are often not found in techniques which use abstract machine

learning algorithms alone. We have argued that these techniques generalize well beyond just phase identification, and have listed properties for which these techniques will be helpful.

CHAPTER 7

BATTERY STORAGE POLICY WITH DEGRADATION MITIGATION

7.1 Decoupled Degradation Valuation and Properties

Reference [109] introduced a linear program (LP) for valuating energy storage systems. The LP partitions the battery's available capacity at each hour $t = 1, 2, ..., \mathcal{T}$ into a set of profitable actions. These actions are discharging, d_t, providing regulation up services, r_t^u, and providing regulation down services, r_t^d. d_t is negative when the battery charges. r_t^u and r_t^d are strictly nonnegative. Each action has a corresponding revenue. These are the locational marginal price, LMP_t, which is the revenue from discharging $1MWh$ of energy, and $Reg_t^{u/d}$, the revenue from committing $1MWh$ of battery capacity to regulation services at hour t. A negative revenue, O_t, is included to incur small profit losses from battery use. It represents the cost of operation and maintenance, but its effect is small.

Only a fraction of the capacity committed to either regulation services will be used in real time operations. The amount used in regulation up is sold at the locational marginal price, and the amount used in regulation down is bought at this price. Denote these proportions as p_t^u and p_t^d for regulation up and down respectively. Then regulation up and down have an additional source of revenue through real time energy exchange given by $LMP_t p_t^{u/d} r_t^{u/d}$.

The decision variables of the LP are formed by appending a state variable, S_t, to the above actions. S_t represents the battery's State of Charge (SoC) at the beginning of hour t and follows simple update dynamics. We will place these decision variables in a vector $\mathbf{x}_t$

and contain the respective revenues in a matrix $\mathbf{A}_t$

$$\mathbf{x}_t = \begin{bmatrix} d_t & r_t^u & r_t^d & S_t \end{bmatrix}^T \tag{7.1}$$

$$\mathbf{A}_t = \begin{bmatrix} LMP_t & Reg_t^u & Reg_t^d & 0 \\ 0 & LMP_t \cdot p_t^u & -LMP_t \cdot p_t^d & 0 \\ 0 & -O_t \cdot p_t^u & -O_t \cdot p_t^d & 0 \end{bmatrix} \tag{7.2}$$

This ensures that the sum of entries of the vector $\mathbf{A}_t\mathbf{x}_t$ yeilds the total revenue at hour t. The horizon of the LP is the estimated battery lifetime. We denote it as $\mathcal{T}$. It should be an overestimate of the battery's true lifetime. Decision variables beyond the true lifetime can be discarded later.

The LP is then given by objective function (7.3) and constraints (7.4 - 7.12).

$$\max_{\mathbf{x_t}} \quad \sum_{t=0}^{\mathcal{T}} \mathbf{1}^T \mathbf{A}_t \mathbf{x}_t \tag{7.3}$$

$$S_{t+1} = S_t(1-\gamma) - (d_t + p_t^u r_t^u - p_t^d r_t^d) \cdot (1\ hr.) \\ - (|d_t| + p_t^u r_t^u + p_t^d r_t^d) \cdot (1\ hr.) \cdot \rho \tag{7.4}$$

$$0 \leq S_t \leq E_{max} \tag{7.5}$$

$$(-d_t + r_t^d) \cdot (1\ hr.) \leq E_{max} - S_t \tag{7.6}$$

$$(d_t + r_t^u) \cdot (1\ hr.) \leq S_t \tag{7.7}$$

$$d_t + p_t^u r_t^u - p_t^d r_t^d \leq P_{max} \tag{7.8}$$

$$-d_t + p_t^d r_t^d - p_t^u r_t^u \leq P_{max} \tag{7.9}$$

$$d_t + r_t^u \leq P_{max} \tag{7.10}$$

$$-d_t + r_t^d \leq P_{max} \tag{7.11}$$

$$r_t^u, r_t^d \geq 0 \tag{7.12}$$

The first constraint is the update equation for the battery's state of charge. The constraint's first term is the battery's self-discharge which occurs with rate γ. The second term is the change in energy from the battery's actions, and the third term represents resistive losses that scale with total output power. The resistive losses, ρ, are derived from the battery's round-trip efficiency κ.

$$\rho = 1 - \sqrt{\kappa}$$

Constraints (7.5), (7.6), and (7.7) capture the fact that the battery's capacity must be partitioned. E_{max} is the battery's maximum state of charge. These constraints ensure that no physical constraint is violated even when all of the committed regulation capacity is used.

Finally, the battery's total output power is constrained by constraints (7.8), (7.9), (7.10), and (7.11). P_{max} is the battery's maximum power output.

Degradation is implemented by partitioning the LP into segments of T hours and optimizing each segment sequentially. The battery's degradation over each segment is

calculated at the end of each iteration. From this calculation, a new value of E_{max} is fed into the next segment's constraints.

T determines a trade-off between optimality and degradation accuracy. Higher values of T will yield more optimal decision variables, but lower values of T (and thus more degradation updates) will lead to more accurate values of E_{max}. T does not need to be very small, however, because E_{max} changes rather slowly with time ($\approxeq 3\%$ per year). Furthermore, T does not need to be too large because the final outputs of the LP, as a function of T, converge when T exceeds 2 months. In this dissertation, T was chosen to represent yearly segments with an expected lifetime of 15 years ($N = 15, T = 8760$ hours).

In this optimization scheme and all optimization schemes to come, we considered a realization of prices that combines the methods of future price curve modeling in [109] with expert opinions and industry price models. The two data sets used (one for locational marginal prices and one for ancillary service prices) can be found at [77].

The External Degradation LP relies on long term forecasting of market prices. These prices are quite volatile in practice, however, so the results of the LP represent a clairvoyant upper bound on the actual value of a BESS.

The constraints of the external degradation LP imply the existence of a region $\mathcal{I} \subset (0, Q_{max})$ in which the state of charge will tend to reside.

Lemma 22. *The maximum value of $r_t^u + r_t^d$ is $E_{max}/(2\ hrs.) + P_{max}$. This maximum is achievable iff. $S_t \in \mathcal{I} = [u, v]$ where u and v are the minimum and maximum of $P_{max} \cdot (1\ hr.)$, $E_{max} - P_{max} \cdot (1\ hr.)$ respectively.*

Proof.

Summing (7.6) and (7.11) yields (7.13). Summing (7.7) and (7.10) yields (7.14)

$$r_t^d + r_t^u \leq E_{max}/(1\ hr.) - (S_t/(1\ hr.) - P_{max}) \tag{7.13}$$

$$r_t^d + r_t^u \leq S_t/(1\ hr.) + P_{max} \tag{7.14}$$

summing (7.13) and (7.14) and dividing by 2 then yields

$$r_t^d + r_t^u \leq E_{max}/(2\ hrs.) + P_{max} \tag{7.15}$$

which is the proposed bound.

To show that this bound is achievable iff. S_t is in the proposed region, we must consider two cases.

First, suppose $P_{max} \leq E_{max}/(2\ hrs.)$ (the typical case in practice). Then if $S_t < P_{max} \cdot (1\ hr.)$, the right hand side of (7.14) is less than $2P_{max}$ which is less than or equal to $E_{max}/(2\ hrs.) + P_{max}$. Thus $r_t^u + r_t^d$ cannot meet the proposed bound. If, however,

$$S_t > E_{max} - P_{max} \cdot (1\ hr.) \tag{7.16}$$

the right hand side of (7.13) again becomes less than $2P_{max}$. Thus the bound (7.15) cannot be achieved outside of $\mathcal{I}$. Within $\mathcal{I}$, however, the right hand sides of both (7.13) and (7.14) are larger than the proposed bound.

Similar logic holds for the case where $P_{max} > E_{max}/(2\ hrs.)$. If $S_t > P_{max} \cdot (1\ hr.)$, the right hand side of (7.13) becomes E_{max} which is less than $E_{max}/(2\ hrs.) + P_{max}$, and if $S_t < E_{max} - P_{max} \cdot (1\ hr.)$, the right hand side of (7.14) becomes E_{max}. Again, both right hand sides are larger than the proposed bound when $S_t \in \mathcal{I}$. □

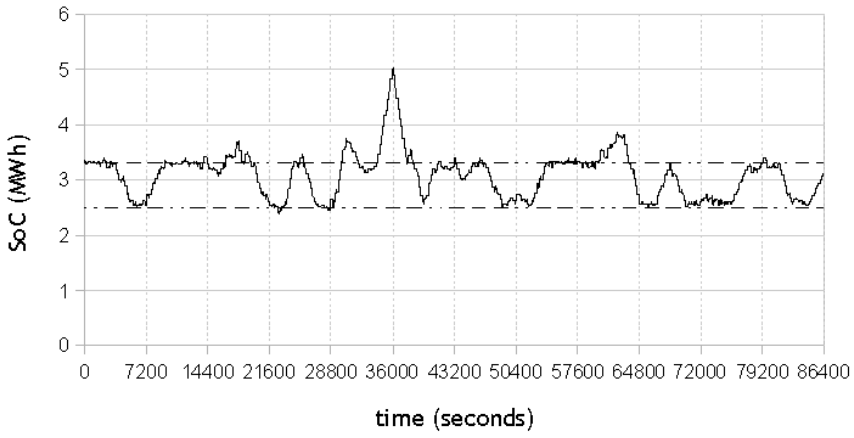

Figure 7.1: Time series data of the SoC over a sample 24 hour window.

The State of Charge will tend to stay in this region because maximizing both regulation variables simultaneously yields high profit. Indeed, we do observe that the SoC is attracted to this region. This is clearly seen in Figure 7.1

Figure 7.1 shows that the State of Charge Profile actually 'bounces' between the upper and lower bounds of $\mathcal{I}$. This can be interpreted as the LP maximizing the arbitrage revenue component while staying in the semi-stable region.

This semi-stable region is the reason that the state space battery model is centered in such a way that the origin of the upper left subsystem corresponds to a SoC of $\frac{1}{2}$.

An approximation to this LP can be formed by introducing a penalty function on deviations of the State of Charge from the semi-stable region and considering only the arbitrage component of the revenue stream (and assuming $r_t^d + r_t^u$ is always maximum).

Specifically, we can form the quadratic cost function:

$$\gamma_c Q_{max}(SoC - \frac{1}{2})^2 = \gamma_c(Q_{SEI} + Q_{ION} - \frac{1}{2}Q_{max})^2 \tag{7.17}$$

$$= \gamma_c(\zeta_{SEI} + \frac{1}{2}\gamma_{SEI} + \zeta_{ION} + \frac{1}{2}\gamma_{ION} - \frac{1}{2}Q_{max})^2 \tag{7.18}$$

$$= \gamma_c(\zeta_{SEI} + \zeta_{ION})^2 \tag{7.19}$$

$$= \gamma_c \zeta^T Q \zeta \tag{7.20}$$

where

$$Q = \gamma_c \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{7.21}$$

This is a positive semidefinite matrix and can thus be easily implemented into various control schemes (with additional costs for charging and discharging in arbitrage).

It is useful to think of the state of charge time series as the sum of a low frequency component and a high frequency perturbation. We will further consider the state of charge over the interval $[t, t+1]$ as a continuous time function $\psi_t(\tau)$, $\tau \in [0, 1](hrs.)$. This should be a decent approximation because the update time for the state of charge time series (4 sec) is much smaller than time scale required for significant degradation ($\approxeq$ 1 year). We will call ψ the state of charge *profile* to distinguish it from the state of charge time series.

The decomposition of ψ will be as follows. Denoting the low frequency component as ψ_M and the high frequency component as ψ_m, we have $\psi_t(\tau) = \psi_{t,M}(\tau) + \psi_{t,m}(\tau)$ where:

$$\psi_{t,M}(\tau) = (S_t - S_{t+1}) \cdot \tau \tag{7.22}$$

$$\psi_{t,m}(\tau) = \psi_t(\tau) - \psi_{t,M}(\tau) \tag{7.23}$$

That is, ψ_M is obtained from linear interpolation of the S decision variables and ψ_m is obtained by removing the low frequency component from ψ. We will consider these continuous functions as valid inputs to the RCA (where, realistically, we run it on the un-approximated time series). We will denote the high frequency cycles as micro cycles and the low frequency ones as macro cycles.

Since the SoC time series frequently jumps in value from $E_{max} \cdot (\frac{1}{2} - R)$ to $E_{max} \cdot (\frac{1}{2} + R)$ and back, we will have a build up of macro cycles with DoD near $2R$. The micro cycles and most of the remaining macro cycles will have DoDs much lower than this. Thus, with respect to DoD, we obtain a bimodal distribution.

The cycles in either of these peaks are equally likely to occur above or below $SoC = \frac{1}{2}E_{max}$. Thus we have symmetry in that, for any range of DoD values, the number of cycles with mean SoC above this line will be nearly equal to the number of cycles with mean SoC below it.

Both of these properties are illustrated in Figure 7.2 which plots the SoC (normalized to E_{max}) against the DoD (also normalized) for a year's worth of cycles. The figure displays a distinct gap in the DoD direction and rough symmetry about the $SoC = \frac{1}{2}$.

7.2 Degradation Linearization

Most ($\approxeq 99\%$) of the realized cycles have current rate on the order of 0.1 or smaller. This is because micro cycles have small DoD, and macro cycles have long time scales. The current rates of the remaining cycles are too few and not large enough ($\approxeq 0.25$) to significantly

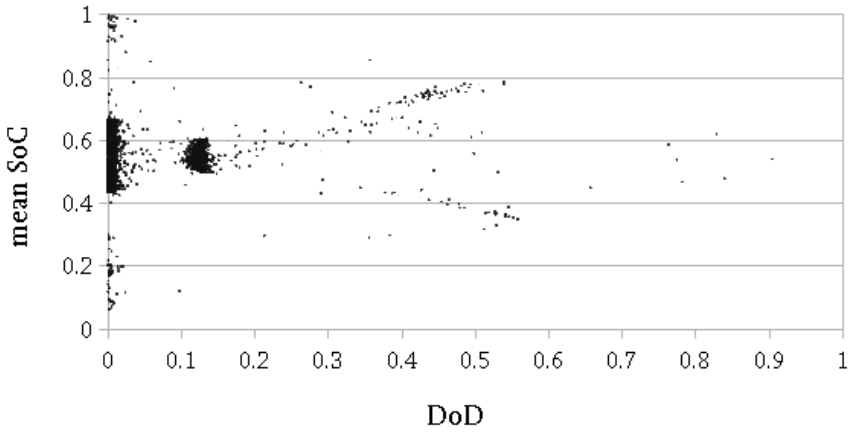

Figure 7.2: Scatterplot of mean normalized SoC vs. normalized DoD.

influence the degradation of the battery. Since the component of degradation due to current rate varies slowly for these small rates, we take f_{CR} to be the constant $J_{CR} = 0.785$ ($CR \approxeq 0.08$).

For the remaining subsections, we will rewrite the cycle degradation component of (2.35) as a Lebesgue integral over a constructed measure space.

Let C_n be the set of all cycles returned from the RCA after iteration n. $\forall c \in C_n$, denote the depth of discharge and mean state of charge of c as $DoD(c)$ and $SoC(c)$ respectively. Let X and Y be unit intervals and let $\mathcal{B}_X$ and $\mathcal{B}_Y$ be the Borel σ-algebras on X and Y. Let ν be the push-forward measure of the counting measure on C_n through the function

$$c \mapsto \begin{bmatrix} DoD(c) \\ SoC(c) \end{bmatrix}$$

Then $(X \times Y, \mathcal{B}_X \otimes \mathcal{B}_Y, \nu)$ is a σ-finite measure space because $\nu(X \times Y) < \infty$. Intuitively, ν counts the number of dots in a given subset of Figure 7.2.

The cycle component of the degradation function can then be written as the following Lebesgue integral.

$$f = J_{CR} \int_{X \times Y} f_{SoC}\, f_{DoD}\, d\nu \tag{7.24}$$

This transfers the sum over each cycle in (2.35) into a sum over possible output values $(f_{CR} \cdot f_{SoC} \cdot f_{DoD})$ times the number of cycles that yield that value $(d\nu)$.

The symmetries discussed so far can be used to reduce the SoC component of the degradation function to a constant (approximately). To see this, let $\pi_x : X \times Y \to X$ be a projection to the x-axis. Let μ be the pushforward measure $\mu = \nu \circ \pi^{-1}$. By the disintegration theorem [93], there exists a family of conditional measures $\{\nu_x\}_{x \in X}$ such that $\nu_x(\{x\} \times Y) = 1\ \forall x \in X$ and (7.24) can be rewritten as the following iterated integral.

$$f = J_{CR} \int_X f_{DoD} \left(\int_{\{x\} \times Y} f_{SoC}\, d\nu_x \right) d\mu \tag{7.25}$$

Now split the y-axis at $\frac{1}{2}$ and denote the lower and upper halves as Y^- and Y^+ respectively. Let h reflect $y \in Y^-$ across the $y = \frac{1}{2}$ axis, i.e. $h(y) = 1 - y$. Symmetry implies that the pushforward of ν_x under h is just ν_x, so the inner integral can be written as follows.

$$\begin{aligned} \int_{\{x\} \times Y} f_{SoC}\, d\nu_x &= \int_{\{x\} \times Y^-} f_{SoC}\, d\nu_x + \int_{\{x\} \times Y^+} f_{SoC}\, d\nu_x \\ &= \int_{\{x\} \times Y^+} \left(f_{SoC} \circ h + f_{SoC} \right) d\nu_x \end{aligned} \tag{7.26}$$

But (relying on the chosen value of $SoC_{ref} = \frac{1}{2}$)

$$\begin{aligned} \left(f_{SoC} \circ h + f_{SoC} \right)(y) &= e^{k_{SoC}(1 - y - \frac{1}{2})} + e^{k_{SoC}(y - \frac{1}{2})} \\ &= 2cosh\left(k_{SoC} \left(y - \frac{1}{2} \right) \right) \end{aligned}$$

which varies slowly for $y \in Y^+$. We can therefore approximate it as a constant. We chose to

take its average over Y^+, $J_{SoC} = 1.0422$, as the constant in question. Equation (7.26) then becomes $2J_{SoC} \cdot \nu(Y^+)$. Symmetry implies that Y^+ has ν_x-measure $\frac{1}{2}$ $\forall x \in X$, so this is just J_{SoC}.

The degradation function now takes the form

$$f = J_{CR} \cdot J_{SoC} \int_X f_{DoD} d\mu \tag{7.27}$$

(7.27) is a 1-dimensional analog of (7.24) which relies on DoD alone.

The integrand in (7.27) can be split by cycle type. Explicitly, we create two new measures for $A \in \mathcal{B}_X$, one that counts the number of *micro* cycles with DoD in A and one that counts the number of *macro* cycles with DoD in A. We call these μ_m and μ_M respectively. Both are dominated by μ because a subset cannot contain micro or macro cycles if it does not contain any cycles. Thus there exist Radon-Nikodym derivatives g_m and g_M such that

$$\mu_m(A) = \int_A g_m d\mu \tag{7.28}$$

$$\mu_M(A) = \int_A g_M d\mu \tag{7.29}$$

clearly $\mu(A) = \mu_m(A) + \mu_M(A)$ $\forall$ $A \in \mathcal{B}_X$ because the number of cycles in A is equal to the number of mirco cycles in A plus the number of macro cycles in A. Then (7.27) becomes

$$f = J_{CR} J_{SoC} \cdot \big(\int_X f_{DoD} \, d\mu_m + \int_X f_{DoD} \, d\mu_M \big) \tag{7.30}$$

All micro cycles are contained in the first peak of the bimodal DoD distribution.

On the other hand, macro cycles can belong to either peak. Since the slope of f_{DoD} is different at both of these peaks, it may be thought that no one line can approximate the degradation from a set of cycles belonging to both. However, there are so few cycles in between these peaks that a line from zero to the second peak will actually work as a decent

approximation. This, of course, requires that we know where the second peak is; luckily, we already know that this is $2R$. We can thus fully linearize (7.30) as

$$f = J_{CR} \cdot J_{SoC} \cdot \big(a_1 \int_X x \, d\mu_m + a_2 \int_X x \, d\mu_M\big) \tag{7.31}$$

Rewrite (7.31) as

$$f = J_{CR} J_{SoC} \big(\int_X x \, (a_1 g_m + a_2 g_M) \; d\mu\big) \tag{7.32}$$

and define $g = a_1 g_m + a_2 g_M$. This is the Raydon-Nikodym derivative of a new measure that counts a scaled version of the macro cycles and adds to it a scaled version of the micro cycles. If a_1 and a_2 are of the same magnitude, then this new measure can be interpreted as an approximation to the measure that counts the number of cycles in X of an augmented state of charge profile $\tilde{\psi}_t(\tau) = a_1 \psi_{t,m}(\tau) + a_2 \psi_{t,M}(\tau)$.

Under this interpretation, the integrand in (7.32) is equivalent to the total depth of discharge traversed by $\tilde{\psi}_t$. This is equal to half of the total absolute change in $\tilde{\psi}_t$. We can thus model the integral in (7.32) as the nonlinear functional

$$\phi(\tilde{\psi}_t) = \frac{1}{2} \int_0^1 \left| \frac{d\tilde{\psi}_t(\tau)}{d\tau} \right| d\tau. \tag{7.33}$$

We will drop the subscript t on all ψ and $\tilde{\psi}$ symbols. For what remains of this subsection, all calculations are considered within hour t.

Considering $a_1 \psi_m$ as a small perturbation on $a_2 \psi_M$, we can approximate (7.33) by the functional Taylor series [31, Appendix A]

$$\phi(\tilde{\psi}) \approx \phi(a_2 \psi_M) + a_1 \frac{d\phi(a_2 \psi_M + \varepsilon \psi_m)}{d\varepsilon}|_{\varepsilon=0} + \frac{1}{2} {a_1}^2 \frac{d^2 \phi(a_2 \psi_M + \varepsilon \psi_m)}{d\varepsilon^2}|_{\varepsilon=0}$$

The zero order term is just

$$\frac{1}{2}|a_2\psi'_M| = \frac{a_2}{2}\big|(1-\rho)(p_t^d r_t^d) - (1+\rho)(p_t^u r_t^u) - d_t - \rho|d_t|\big|$$

as no term in ψ'_M has intra-hour time dependence.

To obtain the first order term, we must differentiate the non-differentiable integrand of (7.33). Thus we approximate this integrand with the differentiable function $((\frac{d\tilde{\psi}}{dt})^2 + u^2)^{\frac{1}{2}}$ and consider what happens as $u \to 0$. The first order term is then

$$\frac{a_1 a_2 \psi'_M}{2(a_2^2\psi'^2_M + u^2)^{\frac{1}{2}}} \int_0^1 \psi'_m dt \tag{7.34}$$

This term will vanish because within the hour,

$$\psi'_m(t) = r_t^u(f_t^u - p_t^u) + r_t^d(p_t^d - f_t^d) \tag{7.35}$$

(where the loss factor ρ has been ignored because all terms involving it in ψ'_m are small) and $\int_0^1 f_t^{u/d} dt = p_t^{u/d}$.

The second order term is given by

$$\frac{a_1^2 u^2}{4(a_2^2\psi'^2_M + u^2)^{\frac{3}{2}}} \int_0^1 \psi'^2_m dt \tag{7.36}$$

But

$$\int_0^1 \psi'^2_M dt = \mathbb{V}ar\left[r_t^u f^u(t)\right] + \mathbb{V}ar\left[r_t^d f^d(t)\right] - \mathbb{E}\left[2r_t^u r_t^d \tilde{f}^u(t)\tilde{f}^d(t))\right] \tag{7.37}$$

where $\tilde{f}^u(t) = (f^u(t) - p_t^u)$, $\tilde{f}^d(t) = (f^d(t) - p_t^d)$, and the statistics are taken as time integrations over the hour under consideration. We will write $\mathbb{V}ar\left[f^u(t)\right]$ as $\sigma^2_{t,u}$ and

$\mathbb{V}ar\left[f^d(t)\right]$ as $\sigma^2_{t,d}$. These can be calculated (or estimated if the signal is not known [41]) before simulation.

Since f^u and f^d are never simultaneously nonzero and $\mathbb{E}[f^{u/d}] = p_t^{u/d}$, the expectation term can be calculated as $-2r_t^u r_t^d p_t^u p_t^d$. The variance terms can be written as

$$(r_t^u + r_t^d)(r_t^u \sigma^2_{t,u} + r_t^d \sigma^2_{t,d}) - r_t^u r_t^d (\sigma^2_{t,u} + \sigma^2_{t,d})$$

Finally, $\frac{u^2}{(a_2^2 \psi'^2_M + u^2)^{\frac{3}{2}}} \to \frac{2}{a_2}\delta(\psi'_M)$ as $u \to 0$, where $\delta(\cdot)$ is the dirac-delta impulse function. Thus the second order term is nonzero only during hours in which $\psi'_M = 0$. This implies that, to the second order, there is complete separation between the macro and micro cycles in terms of degradation.

The second order term is then:

$$\delta(\psi'_M)\frac{a_1^2}{2a_2}\Big((r_t^u + r_t^d)(r_t^u \sigma^2_{t,u} + r_t^d \sigma^2_{t,d}) - r_t^u r_t^d (\sigma^2_{t,u} + \sigma^2_{t,d} + 2p_t^u p_t^d)\Big) \tag{7.38}$$

For a fully linear model, we must find a way to remove the absolute value and delta functions from these formulas. We must also remove the quadratic regulation terms. We choose to do this statistically, i.e. by considering the effects of summing each hour's degradation component over the time interval simulated (e.g. one year).

Let U_z be the set of hours in which $\psi'_M = 0$. Let H be the set of all hours simulated in a given block and $p_z = \frac{|U_z|}{|H|}$. Unfortunately, this must be known before the simulation. However, if it is guessed and updated, it only takes about two iterations for it to converge.

For hours in U_z, there is no change in state of charge to compute. In $H - U_z$, we assume the state of charge increases and decreases with equal probability and assume further that d_t is positive when it decreases and negative otherwise. The absolute value function in the zero order term can then be approximated as $\frac{1}{2}(1 - p_z)|d_t|$ and since d_t appears directly in

the objective function, it can be split into two variables, d'_t and c'_t (for discharge and charge) such that $d_t = d'_t - c'_t$ and $|d_t| = d'_t + c'_t$.

We evaluate the second order term similarly by replacing $\delta(\psi'_M)$ with it's expectation over several possible ψ'_M (as many will occur over the course of a year), p_z.

Finally, the constraints of the LP enforce that $(r_t^u + r_t^d) \leq 2P_{max}$ in $\mathcal{I}$. In fact, maximizing this term is the reason that the SoC time series stays in this region in the first place. Thus we replace the sum with this bound entirely. We also replace the product $r_t^d r_t^u$ with $\frac{1}{2}P_{max}(r_t^u + r_t^d)$ by splitting the term into $\frac{1}{2}(r_t^u r_t^d + r_t^u r_t^d)$ and setting $r_t^u = P_{max}$ in the first term and $r_t^d = P_{max}$ in the second. We do not further approximate this sum as $2P_{max}$ because it is already linear.

We summarize the simplified model in (7.39 - 7.42)

$$\mathbf{b} = \begin{bmatrix} \frac{a_2}{4}(1 - p_z) \\ \frac{a_2}{4}(1 - p_z) \\ \frac{a_1^2}{2a_2} p_z P_{max}(1.5\sigma_{t,u}^2 - 0.5\sigma_{t,d}^2 - p_t^u p_t^d) \\ \frac{a_1^2}{2a_2} p_z P_{max}(1.5\sigma_{t,d}^2 - 0.5\sigma_{t,u}^2 - p_t^u p_t^d) \\ 0 \end{bmatrix} \tag{7.39}$$

$$X_t = \mathbf{b}^T \begin{bmatrix} c'_t \\ d'_t \\ r_t^u \\ r_t^d \\ S_t \end{bmatrix} \tag{7.40}$$

$$deg_n = \sum_{t=0}^{T} X_t + k_t \cdot T \tag{7.41}$$

$$E_{max}^{(n+1)} = r_1 e^{-r_2 \sum_{\eta=1}^{n} deg_\eta} + (1 - r_1) e^{\sum_{\eta=1}^{n} deg_\eta} \tag{7.42}$$

where $a_1 = \frac{2R}{f_{DoD}(2R)}$, $a_2 = 1 \times 10^{-4}$, and k is a tuning parameter defined in the next section.

The final two calculations remain external to the optimization problem. However, the re-simulation step and the RCA algorithm are eliminated from the calculation. Most importantly, however, is that this degradation is explicitly a function on the decision variables of the LP, and can be incorporated *internal* to the optimization.

To test this model, a linear term is added to the objective function, (7.3), to punish high degradation. This leads to a modified version of the original linear program which we will call *Internal Degredation* to contrast it with the external procedure.

First, because we have split the discharge variable into its positive and negative parts (d'_t discharging and c'_t for charging), we need to modify the decision vector $\mathbf{x}_t$ and the revenue matrix $\mathbf{A}_t$.

$$\mathbf{x}'_t = \begin{bmatrix} c'_t & d'_t & r^u_t & r^d_t & S_t \end{bmatrix}^T \tag{7.43}$$

$$\mathbf{A}'_t = \begin{bmatrix} -LMP_t & LMP_t & LMP_t \cdot p^u_t & -LMP_t \cdot p^d_t & 0 \\ 0 & 0 & Reg^u_t & Reg^d_t & 0 \\ 0 & 0 & -O_t \cdot p^u_t & -O_t \cdot p^d_t & 0 \end{bmatrix} \tag{7.44}$$

The optimization problem then takes the form.

$$\max_{\mathbf{x}'_t} \sum_{t=0}^{\mathcal{T}} (\mathbf{1}^T \mathbf{A}'_t \mathbf{x}'_t + M^{deg}_t X_t) \tag{7.45}$$

Subject to

$$S_{t+1} = S_t(1-\gamma) - ({d'}_t - {c'}_t + p_t^u r_t^u - p_t^d r_t^d) \cdot (1\ hr.) \\ - ({d'}_t + {c'}_t + p_t^u r_t^u + p_t^d r_t^d) \cdot (1\ hr.)\rho \tag{7.46}$$

$$0 \ \leq S_t \ \leq \ E_{max} \tag{7.47}$$

$$({c'}_t + r_t^d) \cdot (1\ hr.) \ \leq \ E_{max} - S_t \tag{7.48}$$

$$({d'}_t + r_t^u) \cdot (1\ hr.) \ \leq \ S_t \tag{7.49}$$

$${d'}_t + p_t^u r_t^u - p_t^d r_t^d \leq P_{max} \tag{7.50}$$

$${c'}_t + p_t^d r_t^d - p_t^u r_t^u \leq P_{max} \tag{7.51}$$

$${d'}_t + r_t^u \leq P_{max} \tag{7.52}$$

$${c'}_t + r_t^d \leq P_{max} \tag{7.53}$$

$$X_t = \mathbf{b}^T \mathbf{x}'_t \tag{7.54}$$

$$c'_t, d'_t, r_t^u, r_t^d \geq 0 \tag{7.55}$$

The optimization is still split into N iterations of horizon T. At the end of each iteration, (7.41) and (7.42) are calculated and the new E_{max} is fed into the next iteration.

M_t^{deg} introduces a new trade off. If M_t^{deg} is too large, the LP will sacrifice far too much profit in return for extended battery lifetime. Even further, it will disturb the properties used to derive the linear model. Then the term $M_t^{deg} X_t$ may hurt profits without improving lifetime. If M_t^{deg} is too small, however, then the results of the new optimization scheme will reduce to that of the external degradation scheme. We name this procedure *Internal Degredation* to contrast it with the external procedure.

Since a cycle's degradation is on the order of 10^{-6}, and the remaining multipliers in the objective function are on the order of 1 dollar, we should let M_t^{deg} be on the order of 10^6. Several values of M_t^{deg} were tested. The results of these will be presented in the next

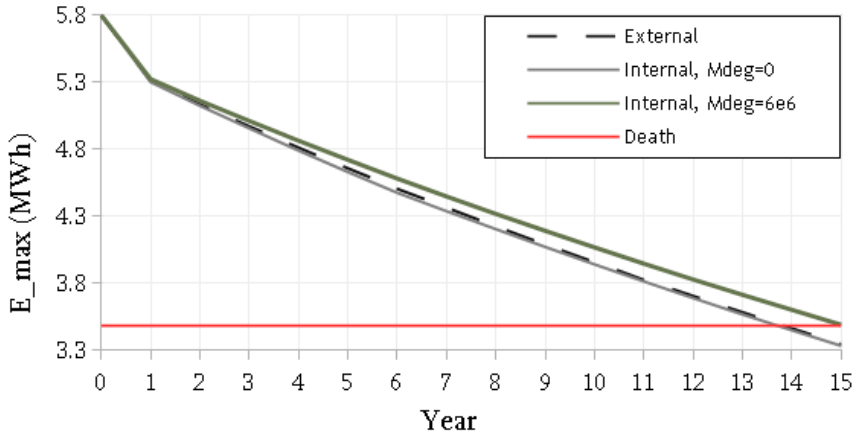

Figure 7.3: E_{max} at the beginning of each year for three optimization schemes.

section.

7.3 Results

The validity of three claims will be shown in this section. The first claim is that the simplification of the degradation function is a good approximation of the actual degradation process. The second is that realistic degradation needs to be considered when valuing a battery energy storage system, and the final claim is that a significant portion of the value lost from the degradation processes can be recovered by using the above internal degradation formulation.

The battery and economic parameters used in the following simulations are shown in Table 7.1.

7.3.1 Linearization Performance

Figure 7.3 shows three plots of E_{max} vs. year number. The lower two of these curves illustrate a direct comparison between the linearized degradation model (estimated) and the

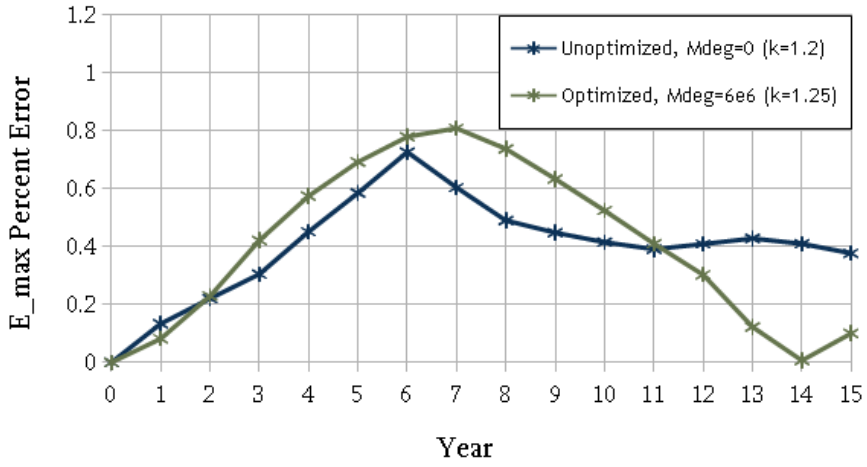

Figure 7.4: Percent error in E_{max} after each year.

Table 7.1: Battery size and economic parameters.

Sizing and Economics			
Parameter	Value	Parameter	Value
E_{max}	$5.8MWh$	P_{max}	$2.53MW$
κ	88%	γ	$1.65\%/mo.$
Energy Investment	$\$614/kWh$	Power Investment	$\$551/kW$
Auxiliary Load	0.875%	Discount Rate	6%

nonlinear, non-closed form degradation model (external) when both are used in the original LP with the same path of prices (or equivalently, when $M_t^{deg} = 0$). On the whole, the curve from the approximated model is lower than that of the nonlinear model. However, the curves are close. Thus the linear model well approximates the more complex one in this case.

Testing that this approximation performs well even when $M_t^{deg} \neq 0$ is critical. To test this, Internal Degradation was run with $M_t^{deg} = 6e6$ and a list of yearly E_{max} values was returned. A hybrid procedure of Internal and External Degradation was then created. This hybrid had the optimization scheme of Internal Degradation (with $M_t^{deg} = 6e6$), but calculated the actual degradation from rainflow counting and the nonlinear model. The

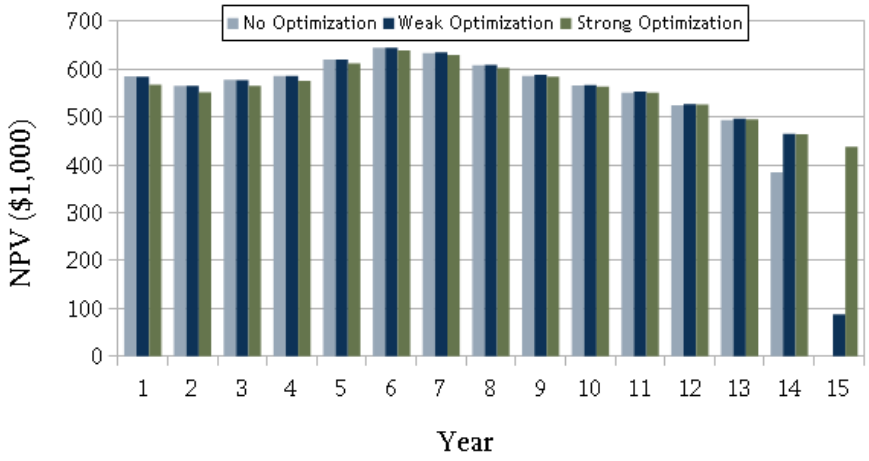

Figure 7.5: Net Present Value (NPV) of each year.

hybrid also returned a list of yearly E_{max} values. The percent differences in the returned E_{max} values are plotted in Figure 7.4 (optimized, $M^{deg} = 6e6$). Figure 7.4 also plots the percent differences in E_{max} for the unoptimized ($M^{deg} = 0$) case. The error in the optimized case is not much worse than that of the unoptimized. However, in the latter case, k was raised slightly. The necessity of raising k implies that the LP properties were less ideal when $M^{deg} \neq 0$. In any case, all errors are bounded by 1%.

7.3.2 Value Loss from Degradation

In a test case, External Degradation was modified to exclude cycle degradation. That is, (2.35) was changed to

$$deg_n = k_t T \tag{7.56}$$

The results of this were compared to an unmodified External Degradation procedure. In the test case, a BESS value of $\$1,700,000$ is obtained. In the latter, this value drops to $\$1,205,000$. Thus $\$495,000$ of value is lost to cycle degradation, a 29.1% loss. Cycle degradation represents a considerable loss to BESS value and needs to be considered.

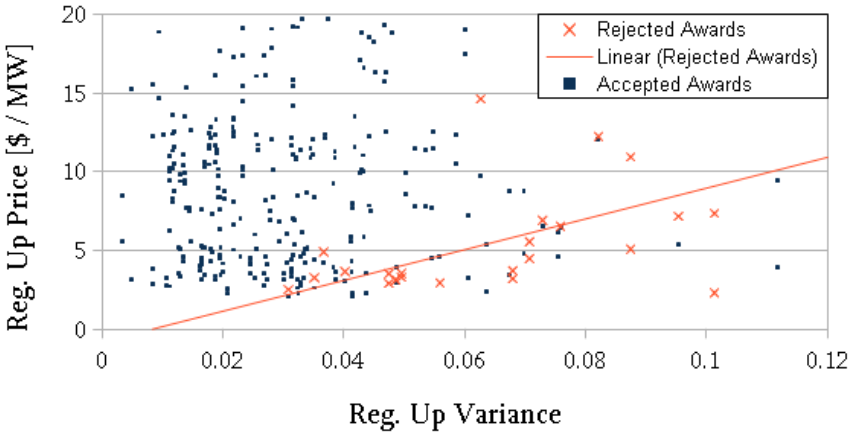

Figure 7.6: Regulation Up signal variance vs. price. The trend line is $y = 97.86x - 0.81$.

7.3.3 Value Recovery

The higher E_{max} curve of Figure 7.3 is found from Internal Degradation with $M_t^{deg} = 6e6$. Critically, this curve crosses the $E_{max} = 3.48$ (60% nominal) line a whole year later than the lower two curves which represent External Degradation. Taking this E_{max} line as the death of the battery, we see that the battery optimized with Internal Degradation provides an additional year of value over the battery optimized with External Degradation.

This is illustrated in Figure 7.5. The two righter bars of this figure (at each year) represent Internal Degradation. The rightmost bar is obtained by using $M^{deg} = 6e6$, and the middle bar is obtained with $M^{deg} = 1e6$. The leftmost bar represents External Degradation. The initial investment is the same in all three cases ($\$6,730,000$, based on battery size). It is observed that Internal Degradation under-performs slightly in the earlier years, but by doing so it gains a large amount of value in the later years. In the strongest modification ($M^{deg} = 6e6$), the battery survives 15 years and the cumulative Net Present Value (NPV) is $\$1,644,000$. This is an increase of $\$439,000$ in NPV over External Degradation, i.e. an 88.7% recovery of the NPV lost ($\$495,000$) from cycle degradation.

With $M_{deg} = 1 \times 10^6$ the battery only survived partially into its 15^{th} year and only recovers 37.5% of the value lost from cycle degradation. Higher M_{deg} may result in large profits by allowing the battery to survive the 16^{th} year, but our pricing data only lasted 15 years so this was not tested.

We can also characterize this improvement through the investment's internal return rate of return (IRR). For External Degradation, the IRR is 8.5%. For Internal Degradation, this raises to 9.15%.

In a more economically realistic analysis, M_{deg} should be chosen with consideration to the cost of the BESS. For example, if the BESS has no cost (and has no cost to set up), then $M_{deg} = 0$ would be appropriate because the battery could be replaced for free. M_{deg} higher than $6e6$ is likely appropriate. This would yield a larger increase in IRR given the realistic initial investment.

It should be noted that the death cutoff will vary by battery. Nonetheless, death will still occur later in the optimized case and value will still be received. For example, if death is taken at 70% nominal capacity, we still obtain nearly an additional year of value with Internal Degradation. In general, this value decreases as the death cutoff increases.

To get a sense of how realistic these results are in practice (since we have assumed perfect forecasting), note that all three scenarios (internal and external degradation) compared assumed perfect price forecasting. Furthermore, all scenarios compared assumed perfect knowledge of the expected value of the regulation signal. Thus we believe that the largest assumption to worry about is that of perfect knowledge of the variance of the regulation signal. Though fair estimates are possible in practice, less accurate signal variances will lead to errors in the degradation model. This will make the internal degradation LP "think" that it will degrade the battery with a slightly higher or lower rate than it actually does. However, since perfect forecasting yields errors of this kind on the order of 1e-5 (each hour), we believe that slight increases in this error will not lead to significantly worse results.

7.3.4 Heuristics and Parameter Estimation

A heuristic for SoH preservation can be inferred from the decision variables chosen by the modified LP. The primary observation is that many potential regulation awards have been rejected completely. As illustrated in Figure 7.6, this occurs when the variance of the regulation signal is high relative to the profit of providing the corresponding service. The other regulation variable is not cut simultaneously, so a small change in SoC occurs during these hours. These differences are compensated by a short series of small charge/discharge variables which take the SoC back to the boundaries of $\mathcal{I}$.

In order to use the approximation of cycle degradation with accuracy, the DoD function parameters must be known accurately. Instead of finding these experimentally, the linear approximation itself can be used to estimate a_1, a_2, and k simultaneously by measuring the SoH of the battery and updating them from this measurement. This will also be considered in future work.

7.3.5 Optimal R

[109] found that, without degradation considerations, the most valuable power to energy ratio, $P_{max} \cdot (1\ hr.)/E_{max}$ (nominal) for wholesale market is $\frac{1}{2}$ because of the increased cost of BESS with higher ratios (i.e. more state of the art in terms of this ratio). The gap jumping property further establishes this optimal ratio. Larger $P_{max} \cdot (1\ hr.)/E_{max}$ will mean that the initial R is larger. With larger R, larger macro cycles will occur, and therefore more degradation. Thus the increased investment in increasing $P_{max} \cdot (1\ hr.)/E_{max}$ will decay more quickly than the investment up to just $P_{max} \cdot (1\ hr.)/E_{max} = \frac{1}{2}$.

7.4 Chapter Summary

This chapter developed a framework for Battery Energy Storage Valuation that is co-optimized with a realistic degradation model. First, BESS optimization was described in detail and a procedure for calculating degradation external to the optimization was explained. It was shown that cycle degradation incurs a 29.1% loss in battery value compared to estimates that did not include it. Properties of the optimized output decisions were then analyzed and a possible heuristic optimization program was discussed. A linear approximation to the degradation function was developed from these properties. By placing the linear model internal to the optimization problem, an additional year of battery lifetime was obtained. This extended lifetime recovered more than 85% of the lost value, reducing the value lost by cycle degradation to just 3.3%. Heuristics for degradation reduction were inferred from the output decision variables of the internal optimization procedure, and an optimal power ratio was argued from the stance of degradation.

CHAPTER 8
CONCLUSION

We have new bounds on information losses from finite data. This began in the form of a relationship between these losses, the expected total variation of the neural model, and the information held in the hidden representation of the feature space. Then, by showing that the total variation term drops quickly with sample size, we obtained bounds that are much tighter and less sensitive to $I(X;Z)$ than previous theory. We provided several applications of this theoretical framework, including an argument for using this research in the development of active learning strategies, an explanation of relevant contradictory experimental work that previously went unexplained, and an application of this theory to low entropy feature space problems. We have provided experiments showing that the bound presented in this chapter are tight to experiment.

We have further provided a novel information theoretic perspective on active learning methods which relies on zero dataset dependent assumptions, and only trivial assumptions overall. we have provided an information theoretic proof of the viability of the facility location function data selection method, and derived a new information theoretic bound which is highly applicable to evaluating other active learning strategies. Experiments show that this bound is very tight, and that it is indicative of dataset quality in terms of classification accuracies.

We have used the theory of information losses to propose the application of two techniques - inverse schur selection and information loading - to the phase identification problem. We observe substatial improvements in phase identification over standard techniques. Furthermore, we have observed that the representations learned upon using these techniques

are much more meaningful than the baseline representations. We have argued that these techniques generalize well beyond just phase identification, and have listed properties of a problem for which these techniques will be helpful.

Finally, we have developed a framework for Battery Energy Storage Valuation that is co-optimized with a realistic degradation model. First, BESS optimization was described in detail and a procedure for calculating degradation in a decoupled way was explained. It was shown that cycle degradation incurs a 29.1% loss in battery value compared to estimates that did not include it. Properties of the optimized output decisions were then analyzed and a possible heuristic optimization program was discussed. A linear approximation to the degradation function was developed from these properties. By placing the linear model internal to the optimization problem, an additional year of battery lifetime was obtained. This extended lifetime recovered more than 85% of the lost value, reducing the value lost by cycle degradation to just 3.3%. Heuristics for degradation reduction were inferred from the output decision variables of the internal optimization procedure, and an optimal power ratio was argued from the stance of degradation.

REFERENCES

[1] and A. Rajeswaran, N. P. Bhatt, and R. Pasumarthy. "A novel approach for phase identification in smart grids using Graph Theory and Principal Component Analysis". In: *2016 American Control Conference (ACC)*. 2016, pp. 5026–5031 (cit. on p. 21).

[2] Alessandro Achille and Stefano Soatto. "Information dropout: Learning optimal representations through noisy computation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018) (cit. on p. 10).

[3] Alessandro Achille and Stefano Soatto. "On the emergence of invariance and disentangling in deep representations". In: *arXiv preprint arXiv:1706.01350* (2017) (cit. on pp. 1, 9, 18, 20).

[4] Max D Anderson and D S Carr. "Battery energy storage technologies". In: *Proceedings of the IEEE* 81.3 (1993), pp. 475–479 (cit. on p. 23).

[5] V. Arya et al. "Phase identification in smart grids". In: *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*. 2011, pp. 25–30 (cit. on p. 21).

[6] Vijay Arya et al. "Systems and methods for phase identification". Pat. US8825416. 2011 (cit. on p. 20).

[7] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. "Agnostic active learning". In: *Journal of Computer and System Sciences* 75.1 (2009), pp. 78–89 (cit. on p. 19).

[8] S. Bashash, S.J. Moura, and H.K. Fathy. "Charge trajectory optimization of plug-in hybrid electric vehicles for energy cost reduction and battery health enhancement". In: *American Control Conference (ACC), 2010* (2010), pp. 5824–5831 (cit. on p. 23).

[9] Ishmael Belghazi et al. "MINE: mutual information neural estimation". In: *arXiv preprint arXiv:1801.04062* (2018) (cit. on pp. 12, 54, 104).

[10] Mikhail Belkin. "Approximation beats concentration? An approximation view on inference with smooth radial kernels". In: *arXiv preprint arXiv:1801.03437* (2018) (cit. on pp. 81, 83).

[11] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011 (cit. on p. 81).

[12] Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. "Importance weighted active learning". In: *arXiv preprint arXiv:0812.4952* (2008) (cit. on p. 19).

[13] Alina Beygelzimer et al. "Agnostic active learning without constraints". In: *Advances in Neural Information Processing Systems*. 2010, pp. 199–207 (cit. on p. 19).

[14] David Blackwell. "Conditional expectation and unbiased sequential estimation". In: *The Annals of Mathematical Statistics* (1947), pp. 105–110 (cit. on p. 15).

[15] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. "Introduction to statistical learning theory". In: *Summer School on Machine Learning*. Springer. 2003, pp. 169–207 (cit. on p. 19).

[16] Kenneth J Caird. *Meter phase identification*. US Patent 8,143,879. 2012 (cit. on pp. 20, 21).

[17] Chao-Shun Chen, Te-Tien Ku, and Chia-Hung Lin. "Design of phase identification system to support three-phase loading balance of distribution feeders". In: *IEEE Transactions on Industry Applications* 48.1 (2012), pp. 191–198 (cit. on pp. 20, 21).

[18] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. "Unsupervised metric learning for face identification in TV video". In: *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE. 2011, pp. 1559–1566 (cit. on p. 76).

[19] David Cohn, Les Atlas, and Richard Ladner. "Improving generalization with active learning". In: *Machine learning* 15.2 (1994), pp. 201–221 (cit. on p. 19).

[20] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012 (cit. on pp. 15, 37).

[21] Imre Csiszar and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011 (cit. on pp. 38, 42).

[22] Tiansong Cui et al. "Optimal co-scheduling of HVAC control and battery management for energy-efficient buildings considering state-of-health degradation". In: *20th Asia and South Pacific Design Automation Conference, ASP-DAC 2015* (2016), pp. 775–780 (cit. on p. 25).

[23] Sanjoy Dasgupta, Daniel J Hsu, and Claire Monteleoni. “A general agnostic active learning algorithm”. In: *Advances in neural information processing systems*. 2008, pp. 353–360 (cit. on p. 19).

[24] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Applications of mathematics. Springer, 1998. ISBN: 9780387984063 (cit. on pp. 40, 43, 47, 52).

[25] M. Dilek, R. P. Broadwater, and R. Sequin. “Phase prediction in distribution systems”. In: *Power Engineering Society Winter Meeting, 2002. IEEE*. Vol. 2. 2002, 985–990 vol.2 (cit. on p. 21).

[26] David L Donoho and Carrie Grimes. “Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data”. In: *Proceedings of the National Academy of Sciences* 100.10 (2003), pp. 5591–5596 (cit. on p. 76).

[27] Monroe D Donsker and SR Srinivasa Varadhan. “Asymptotic evaluation of certain Markov process expectations for large time, I”. In: *Communications on Pure and Applied Mathematics* 28.1 (1975), pp. 1–47 (cit. on p. 52).

[28] Zvi Drezner and Horst W Hamacher. *Facility location: applications and theory*. Springer Science & Business Media, 2001 (cit. on pp. 76, 79).

[29] Madeleine Ecker et al. “Calendar and cycle life study of Li(NiMnCo)O2-based 18650 lithium-ion batteries”. In: *Journal of Power Sources* 248 (2014), pp. 839–851 (cit. on pp. 23, 24).

[30] Ehsan Elhamifar et al. “A convex optimization framework for active learning”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 209–216 (cit. on p. 19).

[31] Eberhard Engel and Reiner M. Dreizler. *Density functional theory: an advanced course*. Vol. 2011. 2011, pp. 499–515. ISBN: 9783642140891 (cit. on p. 125).

[32] Brandon Foggo and Nanpeng Yu. “Interpreting Active Learning Methods Through Information Losses”. In: *arXiv preprint arXiv:1902.09602* (2019) (cit. on p. 95).

[33] Brandon Foggo et al. “Asymptotic Finite Sample Information Losses in Neural Classifiers”. In: *arXiv preprint arXiv:1902.05991* (2019) (cit. on p. 20).

[34] Philipp Fortenbacher, Johanna L. Mathieu, and Goran Andersson. “Modeling, identification, and optimal control of batteries for power system applications”. In: *2014 Power Systems Computation Conference* (2014), pp. 1–7 (cit. on p. 25).

[35] Yoav Freund et al. “Selective sampling using the query by committee algorithm”. In: *Machine learning* 28.2-3 (1997), pp. 133–168 (cit. on p. 20).

[36] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. “Deep bayesian active learning with image data”. In: *arXiv preprint arXiv:1703.02910* (2017) (cit. on p. 19).

[37] Ravi Ganti and Alexander Gray. “Upal: Unbiased pool based active learning”. In: *Artificial Intelligence and Statistics*. 2012, pp. 422–431 (cit. on p. 19).

[38] Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. “Query by committee made real”. In: *Advances in neural information processing systems*. 2006, pp. 443–450 (cit. on p. 20).

[39] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems*. 2014, pp. 2672–2680 (cit. on p. 11).

[40] Yuhong Guo. “Active instance sampling via matrix partition”. In: *Advances in Neural Information Processing Systems*. 2010, pp. 802–810 (cit. on p. 19).

[41] Ryan P Hafen et al. “Requirements for defining utility drive cycles: an exploratory analysis of grid frequency regulation data for establishing battery performance testing standards”. In: (2011) (cit. on p. 127).

[42] Steve Hanneke. *A bound on the label complexity of agnostic active learning*. Citeseer, 2007 (cit. on p. 19).

[43] Moritz Hardt, Benjamin Recht, and Yoram Singer. “Train faster, generalize better: Stability of stochastic gradient descent”. In: *arXiv preprint arXiv:1509.01240* (2015) (cit. on p. 43).

[44] Guannan He et al. “Optimal bidding strategy of battery storage in power markets considering performance-based regulation and battery cycle life”. In: *IEEE Transactions on Smart Grid* (2016), pp. 2359–2367 (cit. on p. 25).

[45] Hongwen He, Rui Xiong, and Jinxin Fan. “Evaluation of lithium-ion battery equivalent circuit models for state of charge estimation by an experimental approach”. In: *Energies* 4.4 (2011), pp. 582–598 (cit. on p. 25).

[46] Steven CH Hoi, Rong Jin, and Michael R Lyu. "Large-scale text categorization by batch mode active learning". In: *Proceedings of the 15th international conference on World Wide Web*. ACM. 2006, pp. 633–642 (cit. on p. 19).

[47] Anderson Hoke et al. "Accounting for lithium-ion battery degradation in electric vehicle charging Optimization". In: *IEEE Journal of Emerging and Selected Topics in Power Electronics* 2.3 (2014), pp. 691–700 (cit. on p. 24).

[48] Frank den Hollander. "Probability theory: The coupling method". In: *Leiden University, Lectures Notes-Mathematical Institute* (2012) (cit. on p. 31).

[49] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5 (1989), pp. 359–366 (cit. on p. 53).

[50] Georgianne Huff et al. "DOE/EPRI 2013 electricity storage handbook in collaboration with NRECA". In: *Report SAND2013* (2013), p. 340 (cit. on p. 7).

[51] Jiayan Jiang, Bo Wang, and Zhuowen Tu. "Unsupervised metric learning by self-smoothing operator". In: (2011) (cit. on p. 76).

[52] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. "Multi-class active learning for image classification". In: (2009) (cit. on p. 19).

[53] Diederik P Kingma, Tim Salimans, and Max Welling. "Variational dropout and the local reparameterization trick". In: *Advances in Neural Information Processing Systems*. 2015, pp. 2575–2583 (cit. on p. 10).

[54] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013) (cit. on p. 10).

[55] Diederik P Kingma et al. "Improved variational inference with inverse autoregressive flow". In: *Advances in Neural Information Processing Systems*. 2016, pp. 4743–4751 (cit. on p. 10).

[56] Mark A Kon and Louise A Raphael. "Approximating functions in reproducing kernel Hilbert spaces via statistical learning theory". In: *Wavelets and Splines: Athens 2005* (2006), pp. 271–286 (cit. on p. 85).

[57] Igor Kononenko. "Bayesian neural networks". In: *Biological Cybernetics* 61.5 (1989), pp. 361–370 (cit. on p. 10).

[58] Andreas Krause, Ajit Singh, and Carlos Guestrin. “Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies”. In: *Journal of Machine Learning Research* 9.Feb (2008), pp. 235–284 (cit. on p. 19).

[59] Jan Kremer, Kim Steenstrup Pedersen, and Christian Igel. “Active learning with support vector machines”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4.4 (2014), pp. 313–326 (cit. on p. 19).

[60] Richard K Lam, Duc Hoai Tran, and Hen-geul Yeh. “Economics of residential energy arbitrage in California using a PV system with directly connected energy storage”. In: *Green Energy and Systems Conference*. 2015, pp. 67–70. ISBN: 9781467372633 (cit. on p. 23).

[61] JG Liao and Arthur Berg. “Sharpening Jensen’s Inequality”. In: *The American Statistician* (2018), pp. 1–4 (cit. on p. 46).

[62] Yizheng Liao et al. “Unbalanced Three-Phase Distribution Grid Topology Estimation and Bus Phase Identification”. In: *arXiv preprint arXiv:1809.07192* (2018) (cit. on pp. 22, 110).

[63] Lars Maaløe et al. “Auxiliary deep generative models”. In: *arXiv preprint arXiv:1602.05473* (2016) (cit. on p. 10).

[64] Johanna L Mathieu and Joshua A Taylor. “Controlling nonlinearbatteries for power systems : trading off performance and battery life”. In: *Power Systems Computation Conference* (cit. on p. 25).

[65] Leland McInnes and John Healy. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018) (cit. on p. 76).

[66] Prem Melville and Raymond J Mooney. “Diverse ensembles for active learning”. In: *Proceedings of the twenty-first international conference on Machine learning*. ACM. 2004, p. 74 (cit. on p. 19).

[67] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. “Variational dropout sparsifies deep neural networks”. In: *arXiv preprint arXiv:1701.05369* (2017) (cit. on p. 10).

[68] Valentin Muenzel et al. “A multi-factor battery cycle life prediction methodology for optimal battery management”. In: *Proceedings of the 2015 ACM Sixth International*

Conference on Future Energy Systems. 2015, pp. 57–66. ISBN: 9781450336093 (cit. on p. 24).

[69] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. "Estimating divergence functionals and the likelihood ratio by convex risk minimization". In: *IEEE Transactions on Information Theory* 56.11 (2010), pp. 5847–5861 (cit. on pp. 10, 11).

[70] F. Ni et al. "Phase identification in distribution systems by data mining methods". In: *2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)*. 2017, pp. 1–6 (cit. on p. 22).

[71] A. Nottrott, J. Kleissl, and B. Washom. "Energy dispatch schedule optimization and cost benefit analysis for grid-connected, photovoltaic-battery storage systems". In: *Renewable Energy* 55 (2013), pp. 230–240 (cit. on p. 23).

[72] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. "f-gan: Training generative neural samplers using variational divergence minimization". In: *Advances in Neural Information Processing Systems*. 2016, pp. 271–279 (cit. on p. 10).

[73] Alexandre Oudalov, Daniel Chartouni, and Christian Ohler. "Optimizing a battery energy storage system for primary frequency control". In: *IEEE Transactions on Power Systems* 22.3 (2007), pp. 1259–1266 (cit. on p. 23).

[74] S. J. Pappu et al. "Identifying Topology of Low Voltage Distribution Networks Based on Smart Meter Data". In: *IEEE Transactions on Smart Grid* 9.5 (2018), pp. 5113–5122 (cit. on pp. 22, 110).

[75] H. Pezeshki and P. J. Wolfs. "Consumer phase identification in a three phase unbalanced LV distribution network". In: *2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*. 2012, pp. 1–7 (cit. on pp. 22, 110).

[76] Abbas Rajabi-Ghahnavieh and Amir-Sina Hamedi. "Explicit degradation modeling in optimal lead-acid battery use for photovoltaic systems". In: *IET Generation, Transmission & Distribution* 10.4 (2016), pp. 1098–1106 (cit. on p. 24).

[77] *Replication data for: battery storage valuation with optimal degradation - harvard dataverse* (cit. on p. 117).

[78] Danilo Jimenez Rezende and Shakir Mohamed. "Variational inference with normalizing flows". In: *arXiv preprint arXiv:1505.05770* (2015) (cit. on p. 10).

[79] Oren Rippel and Ryan Prescott Adams. “High-dimensional probability estimation with deep density models”. In: *arXiv preprint arXiv:1302.5125* (2013) (cit. on pp. 54, 105).

[80] Sam T Roweis and Lawrence K Saul. “Nonlinear dimensionality reduction by locally linear embedding”. In: *science* 290.5500 (2000), pp. 2323–2326 (cit. on p. 76).

[81] Avraham Ruderman et al. “Tighter variational representations of f-divergences via restriction to probability measures”. In: *arXiv preprint arXiv:1206.4664* (2012) (cit. on p. 12).

[82] Igal Sason. “Entropy bounds for discrete random variables via maximal coupling”. In: *IEEE Transactions on Information Theory* 59.11 (2013), pp. 7118–7131 (cit. on p. 31).

[83] Andrew Michael Saxe et al. “On the Information Bottleneck Theory of Deep Learning”. In: *International Conference on Learning Representations*. 2018 (cit. on pp. 1, 19).

[84] B. K. Seal and M. F. McGranaghan. “Automatic identification of service phase for electric utility customers”. In: *2011 IEEE Power and Energy Society General Meeting*. 2011, pp. 1–3 (cit. on pp. 22, 110).

[85] Ozan Sener and Silvio Savarese. “Active learning for convolutional neural networks: A core-set approach”. In: (2018) (cit. on pp. 19, 76, 79, 108).

[86] Burr Settles. “Active learning”. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6.1 (2012), pp. 1–114 (cit. on pp. 19, 95).

[87] Burr Settles. *Active learning literature survey*. Tech. rep. University of Wisconsin–Madison, Jan. 1995 (cit. on p. 19).

[88] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. “Query by committee”. In: *Proceedings of the fifth annual workshop on Computational learning theory*. ACM. 1992, pp. 287–294 (cit. on p. 19).

[89] Ohad Shamir, Sivan Sabato, and Naftali Tishby. “Learning and generalization with the information bottleneck”. In: *Theoretical Computer Science* 411.29-30 (2010), pp. 2696–2711 (cit. on pp. 17, 20).

[90] Junyi Shen, Serkan Dusmez, and Alireza Khaligh. “Optimization of sizing and battery cycle life in battery/ultracapacitor hybrid energy storage systems for electric

vehicle applications". In: *Industrial Informatics, IEEE Transactions on* 10.4 (2014), pp. 2112–2121 (cit. on p. 24).

[91] T. A. Short. "Advanced Metering for Phase Identification, Transformer Identification, and Secondary Modeling". In: *IEEE Transactions on Smart Grid* 4.2 (2013), pp. 651–658 (cit. on p. 22).

[92] Ravid Shwartz-Ziv and Naftali Tishby. "Opening the black box of deep neural networks via information". In: *arXiv preprint arXiv:1703.00810* (2017) (cit. on pp. 1, 9, 18, 20).

[93] David Simmons. "Conditional measures and conditional expectation; Rohlin's disintegration theorem". In: *Discrete and Continuous Dynamical Systems* 32.7 (2012), pp. 2565–2582 (cit. on p. 123).

[94] Maurice Sion et al. "On general minimax theorems." In: *Pacific Journal of mathematics* 8.1 (1958), pp. 171–176 (cit. on p. 52).

[95] Joshua B Tenenbaum, Vin De Silva, and John C Langford. "A global geometric framework for nonlinear dimensionality reduction". In: *science* 290.5500 (2000), pp. 2319–2323 (cit. on p. 77).

[96] Naftali Tishby, Fernando C Pereira, and William Bialek. "The information bottleneck method". In: *arXiv preprint physics/0004057* (2000) (cit. on pp. 1, 15, 105).

[97] Naftali Tishby and Noga Zaslavsky. "Deep learning and the information bottleneck principle". In: *Information Theory Workshop (ITW), 2015 IEEE*. IEEE. 2015, pp. 1–5 (cit. on pp. 1, 20).

[98] Simon Tong. *Active learning: theory and applications*. Vol. 1. Stanford University USA, 2001 (cit. on pp. 19, 95).

[99] Dustin Tran, Rajesh Ranganath, and David Blei. "Hierarchical Implicit Models and Likelihood-Free Variational Inference". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5529–5539 (cit. on p. 11).

[100] Joaquin Vanschoren et al. "OpenML: Networked Science in Machine Learning". In: *SIGKDD Explorations* 15.2 (2013), pp. 49–60 (cit. on pp. 55, 91).

[101] S Verdu et al. "Generalizing the Fano inequality". In: *IEEE Transactions on Information Theory* 40.4 (1994), pp. 1247–1251 (cit. on p. 15).

[102] W. Wang et al. “Phase Identification in Electric Power Distribution Systems by Clustering of Smart Meter Data”. In: *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2016, pp. 259–265 (cit. on pp. 22, 110).

[103] Miles HF Wen et al. “Phase identification in distribution networks with micro-synchrophasors”. In: *2015 IEEE Power & Energy Society General Meeting*. IEEE. 2015, pp. 1–5 (cit. on pp. 20, 21).

[104] Di Wu et al. “An energy storage assessment: using optimal control strategies to capture multiple services”. In: *2015 IEEE Power & Energy Society General Meeting* (2015), pp. 1–5 (cit. on p. 23).

[105] Eric P Xing et al. “Distance metric learning with application to clustering with side-information”. In: *Advances in neural information processing systems*. 2003, pp. 521–528 (cit. on p. 76).

[106] B. Xu et al. “Modeling of Lithium-Ion Battery Degradation for Cell Life Assessment”. In: *IEEE Transactions on Smart Grid* PP.99 (2016), pp. 1–1 (cit. on p. 29).

[107] M. Xu, R. Li, and F. Li. “Phase Identification With Incomplete Data”. In: *IEEE Transactions on Smart Grid* 9.4 (2018), pp. 2777–2785 (cit. on pp. 22, 110).

[108] Kai Yu, Jinbo Bi, and Volker Tresp. “Active learning via transductive experimental design”. In: *Proceedings of the 23rd international conference on Machine learning*. ACM. 2006, pp. 1081–1088 (cit. on p. 19).

[109] Nanpeng Yu and Brandon Foggo. *Stochastic valuation of energy storage in wholesale power markets*. submitted to Energy Economics, 2016 (cit. on pp. 23, 114, 117, 136).

[110] Shifei Yuan, Hongjie Wu, and Chengliang Yin. “State of charge estimation using the extended Kalman filter for battery management systems based on the ARX battery model”. In: *Energies* 6.1 (2013), pp. 444–470 (cit. on pp. 26, 27).

[111] Bo Zhao et al. “Operation optimization of standalone microgrids considering lifetime characteristics of battery energy storage system”. In: *IEEE Transactions on Sustainable Energy* 4.4 (2013), pp. 934–943 (cit. on p. 25).